Universidade de Brasília
IE - Instituto de Exatas
Departamento de Estatística

Modelos Lineares Generalizados Duplos e Aplicações

Marcus Vinicius Teixeira Borba

Brasília
2014

Marcus Vinicius Teixeira Borba

Modelos Lineares Generalizados Duplos e Aplicações

Monografia apresentada ao Departamento de Estatística da Universidade de Brasília, como requisito parcial para a obtenção do título de Bacharel em Estatística.

Orientador: Prof. Dr. Afrânio Márcio Corrêa Vieira

Brasília
2014

Borba, Marcus
Modelos Lineares Generalizados Duplos e Aplicações
46 páginas
Monografia - Instituto de Exatas da Universidade de Brasília. Departamento de Estatística.

1. modelo duplo
2. modelos de regressão
3. probabilidade e estatística aplicada

I. Universidade de Brasília. Instituto de Exatas. Departamento de Estatística.

Comissão Julgadora:

Prof. Dr. Jhames Matos Sampaio

Prof. Dra Juliana Betini Fachini

Prof. Dr. Afrânio Márcio Corrêa Vieira
Orientador

Agradecimentos

Todo conhecimento científico é construído passo a passo. Os meus primeiros passos foram dados na participação de um programa de iniciação científica PET do Departamento de Matemática da UnB. Ao longo dos anos e com a ajuda de alguns excelentes professores fomos escalando alguns degraus. Acredito que, pelo menos alguma parte do conhecimento adquirido por um estudante ao longo de vários anos, num determinado ramo do conhecimento científico, só tem utilidade se for bem compartilhado com futuros alunos e colegas de trabalho. Espero que nossa contribuição humilde possa servir de norte aos estudantes que lutam por um lugar ao sol no futuro.

Agradeço ao meu orientador Prof. Afrânio, aos meus colaboradores, aos funcionários do Departamento de Estatística da Universidade de Brasília, aos meus amigos, à minha família, em especial minha mãe, aos professores do Departamento de Estatística, em especial os professores Dr. Jhames Sampaio e Dra. Juliana Fachini, ao Departamento de Matemática da Universidade de Brasília, em especial os professores Dr. Celius Magalhães, Dr. Hemar Godinho, Dra. Cátia Gonçalves e Msc. Lineu Neto. E acima de tudo, agradeço à Jesus nosso Mestre e Deus nosso Pai.

Resumo

Modelos Lineares Generalizados são largamente utilizados para análise de experimentos planejados. Entretanto, na presença de fatores de ruídos que afetam a variabilidade, encontramos resultados indesejáveis na inferência estatística dos dados.

Neste trabalho abordamos uma modelagem simultânea de média e dispersão em experimentos em blocos casualisados. O estudo tem como objetivo verificar a eficiência da modelagem dupla como alternativa ao planejamento robusto bem como a comparação com modelos mais simples. Os resultados reforçam os critérios da parcimônia e separação necessários numa análise robusta de experimentos.

Keywords: modelo linear generalizado duplo; quase-verossimilhança estendida; delineamento robusto; produção agrícola.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A estatística experimental no início do século XX utilizava métodos em que os dados experimentais seriam analisados através de modelos que consideram uma variância residual constante (ou homogênea), como pressuposto inicial. Entretanto, essa pressuposição mostra-se relativamente forte quando se está diante de situações para as quais fatores externos exercem considerável influência nas medidas experimentais. Entre os métodos teóricos abordados nesse assunto ressalta-se o Planejamento Robusto (TAGUCHI(1985)) que, resumidamente, tinha o objetivo de reduzir a variabilidade de processos industriais mantendo a média de uma característica de qualidade em um valor nominal.

Um problema abordado na modelagem clássica daquela época era o aspecto da heterogeneidade inerente aos problemas experimentais, cujo controle local era feito através da aleatorização da alocação de tratamentos dentro de blocos (COCHRAN; COX (1957)). Um outro problema relacionado com a heterogeneidade mas também associado com distribuições probabilísticas discretas como a Binomial e Poisson, abordado naquela época, é denominado “sobredispersão” (ou "superdispersão") onde a variabilidade dos dados observados é bem maior que a variabilidade esperada pelo modelo. Em qualquer caso, os dados observados em um determinado experimento, apresenta-

vam uma variabilidade gerada por fatores não controláveis com possíveis causas dos chamados "ruídos" ou fatores de variação ou mesmo o pressuposto de uma distribuição inadequada na modelagem dos dados.

Com pesquisadores buscando novas formas de se planejar experimentos, surgiu uma nota de revisão (PREGIBON (1984)) que retoma o problema de modelagem dentro do "framework" dos modelos lineares generalizados (MLG). Utilizando a idéia dos MLG, como proposta de resolver algumas destas questões, uma nova extensão dos MLG (SMYTH (1989)) denominada Modelo Linear Generalizado Duplo (MLGD), dos fatores de controle e dos fatores de variação ou ruído (também denominados como efeitos de locação e dispersão, respectivamente). Aplicou-se este modelo para a média e a variância, simultaneamente, com a variância estruturada por meio de um preditor linear, que permite incorporar variáveis externas e fatores de ruído.

Esses e outros problemas motivam a pesquisa de métodos que modelem simultaneamente a média e a dispersão, aplicados à pesquisa experimental. Atualmente, alguns pesquisadores (VIEIRA et al. (2010),VIEIRA et al. (2011)) vem utilizando a modelagem dupla como proposta de solução de problemas experimentais.

Neste trabalho, serão descritos os modelos lineares generalizados (MLG), os algoritmos de estimação, suas extensões e medidas de discrepâncias para diagnóstico de ajuste dos modelos. A seguir descreveremos os algoritmos de estimação dos modelos lineares generalizados duplos (MLGD) que modelam de forma alternada e iterativa a média e a dispersão dos dados. E ao final do trabalho, aplicaremos este modelo a um conjunto de dados, fazendo as devidas comparações e conclusões.

1.1 Objetivos

O objetivo deste trabalho é estudar os fundamentos dos Modelos Lineares Generalizados Duplos (MLGD), seus algoritmos de estimação e testes de hipóteses, assim como

aplicá-los em um estudo experimental na produção de milho, utilizando informações de variáveis físico-químicas do solo.

1.1.1 Objetivos Específicos

A fim de modelar um experimento em blocos casualisados utilizando o conceito de Planejamento Robusto iremos:

- Revisar os modelos lineares generalizados duplos (MLGD);
- Comparar com outros modelos mais simples;
- Utilizar em exemplos com estimativas ;
- Testar o ajuste dos modelos.

Capítulo 2

Conceitos Básicos

2.1 Modelos Lineares Generalizados

Os Modelos Lineares Generalizados (MLG) foram uma teoria de unificação de técnicas estatísticas proposta no artigo de Nelder; Wedderburn (1972) e depois sendo aperfeiçoado em vários livros e artigos, especialmente no livro de Mccullagh; Nelder (1989). Mais especificamente, considerando uma amostra com n observações independentes , $\mathbf{X}$ uma matriz com $p+1$ colunas e $\mathbf{y}$ um vetor de observações amostrado de $\mathbf{Y}$, são definidos os 3 componentes de um MLG:

1. $\mathbf{Y}$ tem distribuição probabilística como membro da **Família Exponencial** de distribuições, com uma função de probabilidade ou função densidade de probabilidade (para variáveis aleatórias discretas e contínuas, respectivamente)

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

   com média

$$E(Y_i) = \mu_i = b'(\theta_i)$$

e variância

$$Var(Y_i) = a_i(\phi)b''(\theta_i) = a_i(\phi)V(\mu_i) = a_i(\phi)V_i$$

sendo θ_i o parâmetro canônico, $a_i(\phi) = \phi/w_i$, ϕ o parâmetro de dispersão, w_i o "peso a priori" e V_i a função de variância dada por $V_i = \frac{d\mu_i}{d\theta_i}$.

Algumas distribuições membros da família exponencial são a Normal, Poisson, Binomal, Gama, Gaussiana Inversa, Binomial Negativa (com parâmetro k fixo), etc.

2. A matriz $\mathbf{X}$ de covariáveis e fatores estão relacionadas no chamado preditor linear na forma

$$\eta_i = x_i^T\beta;$$

3. Uma **Função de Ligação** monótona (inversível) e diferenciável $g(.)$, que liga o preditor linear η à média de $\mathbf{Y}$ onde escrevemos

$$g(\mu_i) = x_i^T\beta \Longrightarrow \mu_i = g^{-1}(x_i^T\beta)$$

2.2 Algoritmo de Estimação dos parâmetros de um MLG

O método de estimação para o vetor de parâmetros β , proposto por Nelder e Wedderburn (DEMÉTRIO (2002)), utiliza o método de máxima verossimilhança. A log-verossimilhança da família exponencial será dada por

$$l(\theta|\mathbf{y}) = \ln[\prod_{i=1}^{n} f(y_i,\theta_i,\phi)] = \sum_{i=1}^{n} l(\theta_i, y_i) =$$

$$= \sum_{i=1}^{n}\{\frac{1}{a(\phi)}[y_i\theta_i - b(\theta_i)] + c(y_i,\phi)\};$$

com j-ésimo vetor score dado pela equação

$$U_j = \frac{\partial l(\theta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{dl_i}{d\theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{d\eta_i}{d\beta_j} =$$

$$= \sum_{i=1}^{n} \frac{1}{a(\phi)}(y_i - \mu_i)\frac{1}{V(\mu_i)}\frac{d\mu_i}{d\eta_i}x_{ij} \qquad j = 1,...,p \tag{2.2}$$

Curiosamente, a equação funcional de estimação do vetor escore utiliza apenas a média $\mu_\mathbf{i}$ e a função de variância $V(\mu_i)$, não sendo necessária toda a informação da forma funcional da família exponencial. Obviamente que necessitamos de métodos numéricos (Newton-Raphson por exemplo) para calcular numericamente as soluções das equações de estimação, por se tratar de equações não lineares.

As equações de estimação (2.2) são solucionadas utilizando o algoritmo iterativo chamado Iterative Reweighted Least Square (IRLS), que equivale ao método scoring de Fisher) cuja m-ésima iteração é descrita abaixo até que se atinja algum critério de convergência:

Passo 1:

$$\eta_i^{(m)} = \sum_{j=1}^{p} x_{ij}\beta_j^m\,; \qquad \mu_i^{(m)} = g^{-1}(\eta_i^{(m)})\,;$$

Passo 2:

$$q_i^m = \eta_i^{(m)} + (y_i + \mu_i^{(m)})g'(\mu_i^{(m)})\,; \qquad Q^{(m)} = [q_1^{(m)},...,q_n^{(m)}]^T$$

$$W_i^{(m)} = \frac{w_i}{V(\mu_i^{(m)})[g'(\mu_i^{(m)})]^2}\,; \qquad \mathbf{W}^{(m)} = diag\{W_i^{(m)}\}\,;$$

Passo 3:

$$\beta^{(m+1)} = (\mathbf{X}^T\mathbf{W}^{(m)}\mathbf{X}^{-1})^{-1}\mathbf{X}^T\mathbf{W}^m\mathbf{Q}^{(m)}\,;$$

Passo 4: Se o ciclo iterativo convergir, $\hat{\beta} = \beta^{(m)}$. Caso contrário, vá para o Passo 1 utilizando $\beta^{(m)}$ na iteração $m+1$.

2.3 Medidas de Discrepância de um MLG

Como medida de discrepância entre o valor obtido na amostra e o valor predito pelo modelo, Nelder e Wedderburn (1972) propuseram a medida de desvio

$$\frac{D}{\phi} = -2\sum_{i=1}^{n}\{l(\hat{\mu}_i,y_i) - l(y_i,y_i)\}$$

sendo $l(\hat{\mu}_i,y_i)$ o logaritmo da função de verossimilhança calculado utilizando y_i e uma estimativa de μ_i e $l(y_i,y_i)$ é o logaritmo da função de verossimilhança calculado utilizando y_i também como uma estimativa de μ_i. Essa medida é chamada *scaled deviance* e

$$D = \sum_{i=1}^{n} d_i^2$$

é a *deviance* residual , sendo d_i^2 o componente de *deviance*, dada por

$$d_i^2 = -2\int_{y}^{\mu_i} \frac{y - u_i}{V(u_i)}du$$

Assumindo $V(\mu_i)$ e μ_i para cada distribuição pertencente à família exponencial na forma canônica (equação (2.1)), obtêm-se diferentes expressões para a *deviance*. O caso mais simples é a distribuição Normal com $V(\mu_i)$ =1 e $\mu_i = \eta_i$. Neste caso D é a soma de quadrados dos resíduos coincidindo assim com um modelo ANOVA já conhecido. Para o cálculo dos resíduos do modelo ajustado:

1 Resíduo componente da deviance:

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

2 Resíduo de Pearson

$$r_{p_i} = \frac{y_i - \hat{\mu}_i}{V(\hat{\mu}_i)}\,.$$

Adicionalmente, se H_0 e H_a são as hipóteses associadas a dois modelos encaixados de dimensões p e q, respectivamente, com $p < q$, então, sob H_0 , a mudança na *scaled deviance* $D(\hat{\mu_a},\hat{\mu_0}) = D(y,\hat{\mu_0}) - D(y,\hat{\mu_a})$ tem distribuição assintótica χ^2_{q-p}.

Em MLG a função desvio é usada para medir a discrepância de um ajuste, e também pode ser usada para comparar modelos com diferentes preditores lineares e/ou funções de ligação. A função desvio não pode, contudo, ser usada para comparar modelos com diferentes funções de variância ou diferentes estruturas de dispersão, que aparecem na modelagem conjunta da média e dispersão.

2.4 Quase-Verossimilhança

Em situações em que nem sempre é possível assumir uma distribuição probabilística conhecida para os dados, Wedderburn (1974) propôs o método da "quase-verossimilhança" (QV), que relaxa a hipótese de que a distribuição pertença à família exponencial desde que se conheça a relação entre a média e a variância. O logaritmo da função de quase verossimilhança Q será definido por

$$\frac{\partial Q(y_i,\mu_i)}{\partial \mu} = \frac{y_i - \mu_i}{V(\mu_i)} \Longrightarrow Q(y_i,\mu_i) = \int^{\mu} \frac{y-u}{V(u)} du\,. \tag{2.3}$$

Quase verossimilhanças permitem dois tipos de extensões de MLG. Na primeira, MLG com $\phi = 1$, fixo, podem ser estendidos para admitirem ϕ variável; por exemplo, os modelos loglineares de Poisson, para os quais $Var(\mathbf{Y}) = \mu$, podem ser expandidos com $Var(\mathbf{Y}) = \phi\mu$ e $\phi > 1$. Na segunda extensão, $V(\mu)$ pode tomar uma forma que não corresponde àquela, própria de um MLG padrão, por exemplo, $V(\mu) = \mu^{\alpha}$, com α variável e $\alpha \neq 0,1,2,3$ (NELDER; LEE (1991)).

A quase verossimilhança tem as mesmas equações de estimação que os MLG, gerando estimativas de máxima quase verossimilhança no lugar de estimativas de máxima verossimilhança; e também produzindo um desvio (*deviance*) e um resíduo de Pearson.

O método QV pode ser utilizado na modelagem de dados de contagem ou proporção sendo que estes tipos de dados estão sujeito à chamada "superdispersão" onde a variação observada é muito maior do que a variação prevista pelos modelos, baseados na distribuição de Poisson e Binomial (HINDE; DEMÉTRIO (1998)).

Porém, a desvantagem do método QV em assumir o parâmetro de dispersão ϕ constante para todas as observações levou a demanda de utilizar métodos de estimação mais gerais.

2.5 Quase-Verossimilhança Estendida

Este modelo proposto por Nelder; Pregibon (1987) considera situações em que modelamos o parâmetro de dispersão dependente de covariáveis sem considera-lo constante. A chamada quase-verossimilhança estendida (QVE) tem função de quase verossilhança proposta por

$$-2Q^{+}(y_i,\mu_i) = \sum_{i=1}^{n}\{\frac{d_i}{\phi_i} + \ln[2\pi\phi_i V(y_i)]\}\,. \tag{2.4}$$

Para as distribuições Normal e Gaussiana Inversa o logaritmo da função de quase-verossimilhança estendida gera funções do logaritmo da verossimilhança exata e boas aproximações para as demais distribuições. A forma Q^{+} é utilizada para estudos assintóticos de estimadores de máxima verossimilhança mas especificamente nos MLG ela é apenas aplicada nas observações individuais. O método QVE é bastante utilizado em problemas onde a variância apresenta estruturas mais complexas com dispersão variante (LEE et al. (2006)).

Pode-se considerar também a variância pertencente a uma família de funções indexadas por um parâmetro desconhecido λ . Uma família, muito útil, é obtida considerando potências de μ : $V_\lambda(\mu) = \mu^\lambda$ (NELDER; PREGIBON (1987)). Os valores mais comuns de λ são: 0,1,2 e 3; os quais correspondem às funções de variâncias associadas com as distribuições Normal, Poisson, Gama e Inversa Gaussiana, respectivamente.

Para um dado ϕ_i , a menos de uma constante, a QVE é a QV para um modelo com a função de variância $V(\mu_i)$. Desta forma, maximizando Q^+ , com respeito ao vetor de parâmetros β , teremos os mesmos estimadores da QV, com pesos $\frac{1}{\phi_i}$, satisfazendo:

$$\frac{\partial Q^+}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi_i V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0\,,$$

para $j = 1,...,p$, onde p é o número de parâmetros no modelo da média. A QVE fornece um desvio padronizado, o qual pode ser usado como uma medida de discrepância (LEE; NELDER (1998)).

Por outro lado, dado μ_i, a QVE se transforma num MLG com distribuição Gama para a variável resposta d_i onde $E(d_i) = \phi_i$ e $Var(d_i) = 2\phi_i^2$. Assim, maximizamos Q^+ com respeito ao vetor de parâmetros γ originando as equações de estimação:

$$\frac{\partial Q^+}{\partial \gamma_j} = \sum_{i=1}^{n} \frac{di - \phi_i}{\phi_i^2} \frac{\partial \phi_i}{\partial \gamma_j} = 0$$

Segundo Nair (1992), a modelagem Gama para a dispersão é uma boa aproximação até mesmo quando $\mathbf{Y}$ não tem distribuição Normal. Observamos que a distribuição para dispersão é Gama exata se $\mathbf{Y}$ tem uma distribuição normal e no modelo da dispersão é comum tomar a função de ligação logarítmica.

2.6 Modelo Linear Generalizado Duplo

No método de Quase-Verossimilhança, onde assumimos $Var(Y) = \phi\mu$ para acomodar a chamada superdispersão, a otimização do vetor de parâmetros β da equação (2.3), geram estimativas iniciais que coincidem com o MLG ordinário. O parâmetro ϕ é estimado igualando a estatística de Pearson χ^2 ou a deviance residual ao número de seus graus de liberdade. Uma vez obtida a estimativa $\tilde{\phi}$, esta é fixada e o MLG ordinário deve ser reajustado para que seja possível fazer inferência sobre os parâmetros. O vetor

β será o mesmo mas o erro padrão estará ajustado em $\sqrt{\tilde{\phi}}$.

Assim, a função de variância escrita numa forma geral $Var(Y) = \phi V(\mu)$ depende basicamente de dois componentes: um dependendo da média $V(\mu)$ e outro independente da média ϕ. Uma extensão natural deste raciocínio é adotar um outro preditor linear ξ , associado à variação não explicada ϕ por meio de uma função de ligação $g_d(.)$. Este modelo proposto por Smyth (1989) e Nelder; Lee (1991) é chamado Modelo Linear Generalizado Duplo (MLGD) onde

$$\begin{cases} \eta_i = g(\mu_i) = x_i^T \beta \\ \xi_i = g_d(\phi_i) = z_i^T \gamma \end{cases} \tag{2.5}$$

sendo que os vetores de covariáveis ou fatores $\mathbf{z_i}$ estão associados aos parâmetros em γ, que influenciam na variabilidade da variável resposta. Para a estimação dos parâmetros de (2.5), adotamos o seguinte processo iterativo :

(i) Partindo de valores iniciais para os parâmetros com γ fixado, estima-se o vetor β através de um MLG ordinário para a variável resposta Y com peso a priori w_i/ϕ_i ;

(ii) Fixando o vetor β , a estimativa de γ é obtida assumindo ϕ como uma variável resposta com distribuição Gama e ajustando um MLG para o preditor linear $\xi_i = z_i^T \gamma$, fixando-se o parâmetro de dispersão igual a 2. Lembrando que, a justificativa para fixar o modelo para a dispersão como Gama é que o desvio tem uma distribuição próxima da distribuição Gama até mesmo quando Y não tem uma distribuição Normal (Nair, 1992). Note que a distribuição para dispersão é Gama exata se Y tem uma distribuição normal.

Estes 2 passos devem ser alternados até que se atinja algum critério de convergência.

Este procedimento comumente chamado **método iterativo para a modelagem conjunta da média e dispersão** é detalhado a seguir

- Modelo para média

  Sejam y_1,...,y_n , n observações independentes da variável resposta Y, x_1,...,x_p as p covariáveis que afetam a média e β_1,...,β_p os parâmetros do modelo e considere o vetor γ fixado. Escreva $\mu^T = (\mu_1,...,\mu_n)$, $\phi^T = (\phi_1,...,\phi_n)$ onde sabemos que $E(Y_i) = \mu_i$ e $Var(Y_i) = \phi_i V(\mu_i)$. Utilizando o algoritmo IRSL calculamos

$$\beta_{(j)} = (\mathbf{X}^T\mathbf{W}_{(j-1)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_{(j-1)}\mathbf{q}_{(j-1)}$$

  com $\mathbf{X}$ sendo a matriz de planejamento da média de dimensão (nxp); $\mathbf{W}_{(j-1)} = Diag\left(w_{(j-1)1},...,w_{(j-1)n}\right)$ a matriz nxn de pesos com as entradas na diagonal calculadas na forma $w_{(j-1)i} = \left(\frac{\partial \mu_{(j-1)i}}{\partial \eta_{(j-1)i}}\right)^2 \frac{1}{V(\mu_{(j-1)i})}$ e $\mathbf{q}^T = (q_{(j-1)1},...,q_{(j-1)n})$ com $q_{(j-1)i} = \eta_{(j-1)i} + \frac{\partial \mu_{(j-1)i}}{\partial \eta_{(j-1)i}}(y_i - \mu_{(j-1)i})$ nos índices $i = 1,...,n$ e iterações $j = 1,2....$

  O processo iterativo em j continua até que algum critério de convergência seja satisfeito, por exemplo $|\beta_{(j)} - \beta(j-1)| < \epsilon$ para um $\epsilon > 0$ dado. De posse do vetor $\hat{\beta}$ calculamos o vetor $\hat{\mu}$ utilizando a inversa da função de ligação com $\mu = g^{-1}(x_i^T\beta)$.

- Modelo para dispersão

  Com o vetor $\hat{\mu}$ calculado na modelagem da média, calculamos o vetor $\mathbf{d}^*$ dado pela fórmula $d_i^* = \frac{d_i}{1-h_i}$ onde $d_i = 2\int_{\mu_i}^{y_i} \frac{y_i - u}{V(u)}du$ e h_i é o i-ésimo elemento da diagonal da matriz

$$\mathbf{H} = \mathbf{W}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}\mathbf{W}^{\frac{1}{2}} \qquad i = 1,...,n.$$

  Calculado o vetor $\mathbf{d}^*$ e os pesos $1-h_i$ e considerando uma distribuição Gama para o modelo da dispersão afetada pelas covariáveis γ_1,...,γ_q, aplicamos novamente o

algoritmo IRSL calculando a iteração do vetor γ na forma

$$\gamma_j = (\mathbf{U}^T\mathbf{V}_{(j-1)}\mathbf{U})^{-1}\mathbf{U}^T\mathbf{V}_{(j-1)}\mathbf{s}_{(j-1)}.$$

Aqui γ_j é um vetor (qx1), $\mathbf{U}$ a matriz de planejamento (nxq), $\mathbf{V}_{(j-1)} = diag(\nu_{(j-1)1},...,\nu_{(j-1)n})$ a matriz de pesos calculados na expressão $\nu_{(j-1)i} = \left(\frac{\partial\phi_{(j-1)i}}{\partial\zeta_{(j-1)i}}\right)^2 \frac{1}{\phi^2_{(j-1)i}}(1 - h_{(j-1)i})$ para a diagonal e $\mathbf{s}^T_{(j-1)} = (s_{(j-1)1},...,s_{(j-1)n})$ um vetor com

$$s_{(j-1)i} = \zeta_{(j-1)i} + \frac{\partial\zeta_{(j-1)i}}{\partial\phi_{(j-1)i}}(d_{(j-1)i} - \phi_{(j-1)i})$$

para $i = 1,...,n$ e $j = 1,2,....$

Da mesma forma como no modelo da média, estimamos o vetor de parâmetros $\hat{\gamma}$ para calcular a média $\hat{\phi}$, utilizando algum critério de convergência, através da função de ligação g_d na relação $\hat{\phi} = g_d^{-1}(\mathbf{U}\hat{\gamma})$. Geralmente utilizamos g_d como a função log.

Com $\hat{\phi}$ atualizado volta-se novamente no modelo da média para estimar um novo vetor $\hat{\mu}$ e iniciamos um processo iterativo alternando-se entre o modelo da média e o modelo da dispersão até que algum critério de convergência seja atingido. Um critério bastante utilizado é dado na forma

$$\frac{|QD_k^{+A} - QD_{k-1}^{+A}|}{QD_k^{+A}} < \epsilon$$

onde $\epsilon > 0$ e $QD^{+A} = \sum_{i=1}^{n} \frac{d_i^*}{\phi_i} + \sum_{i=1}^{n} \ln[2\pi\phi_i V(y_i)]$ é chamado de quase-desvio estendido.

A seleção dos parâmetros e a verificação da qualidade do ajuste é feita normalmente para modelos MLG com estimação QVE. O problema desta abordagem pura é o pouco tratamento do viés das estimativas de γ. Entretanto, melhorias na aproximação da distribuição dos dados e redução do viés nas estimativas é tratado por Smyth; Verbyla (1999), cujo algoritmo foi implementado no ambiente R (R DEVELOPMENT CORE

TEAM, 2007) na biblioteca dglm utilizando a função dglm(). Este algoritmo será utilizado na análise central do conjunto de dados do capítulo 3.

Em MLG, ϕ e μ são medidas de desempenho para o ruído e para a média, respectivamente. A dependência funcional entre a média e a variância é eliminada através de uma escolha apropriada da função de variância $V(\mu)$. As medidas de desempenho são modeladas através de especificações apropriadas para as funções de ligação da média e da dispersão. O objetivo é encontar modelos aditivos mais simples para a média e a dispersão

Box (1988) considera dois critérios para análise de dados, em experimentos para melhoramento da qualidade, como sendo de grande importância, separação e parcimônia. Separação é a eliminação da dependência funcional entre a média e a variância; e parcimônia é a provisão de um modelo aditivo mais simples possível.

Uma covariável de dispersão pode ou não ser a mesma que uma covariável para a média. Com isto, na formulação do MLG, as duas metas: separação e parcimônia, são interpretadas da seguinte forma:

- Separação: escolher a função de variância $V(\mu)$ adequada para a média;
- Parcimônia: escolher corretamente a função de ligação e o preditor linear para os modelos da média e dispersão.

Utilizando a estimação QVE, Nelder e Lee (1991) abordam que a predição dos modelos têm como objetivo minimizar a variabilidade ajustando a média num valor alvo. Para garantir a qualidade da predição é preciso uma escolha conveniente da função de variância $V(\mu)$ e uma escolha de modelos parcimoniosos com funções de ligação apropriadas para a média e dispersão juntamente com um conjunto parcimonioso de covariáveis no preditor linear. Além disso, a qualidade do ajuste deve ser verificada (McCullagh; Nelder, 1989).

Neste trabalho adotaremos a seguinte estratégia de análise (NELDER; LEE (1991)):

Em situações para as quais não há repetições, como em experimentos fatoriais fracionados cruzados com outra estrutura fatorial, para acomodar efeitos de ruídos, utilizaremos a abordagem:

i. deve-se iniciar o processo de seleção de modelos com um preditor linear maximal para a média, contendo os efeitos principais para os fatores de controle e ruído;

ii. fazer uma busca por interações de grande significância entre os fatores de controle e os de ruído. Interações de ordem superior mas com efeito negligenciável podem ser desconsideradas do modelo e utilizadas para o modelo de dispersão. Se nenhum fator de ruído se mostrou significativo no modelo para a média, a análise pode ser realizada como se cada fator de controle fosse repetido, com o número de observações correspondentes ao número de combinações referente ao delineamento experimental dos fatores de ruído;

iii. encontrado um preditor linear inicial para a média, fixá-lo e iniciar a busca de um preditor linear parcimonioso para a dispersão.

iv. fixado o modelo de dispersão, retomar a busca de um modelo parcimonioso para a média da variável resposta estudada.

Para esta estratégias de análise, a verificação da qualidade de ajuste do modelo deve ser realizada, com base na análise de resíduos dos modelos de média e dispersão.

Capítulo 3

Exemplo de experimento agrícola em blocos casualisados

Num experimento químico em blocos casualisados de 4 tipos de tratamentos (híbridos da planta do milho) com 24 observações divididas em 6 blocos de tamanho idêntico, foram medidas 3 variáveis respostas: Altura da Espiga, Produção de Grãos de milho e Número de Espigas codificadas nas palavras AltEspiga,Produçao e Espigas respectivamente. Neste trabalho analisaremos a variável Produção de Grãos Kg/Ha (Distribuição Normal) levando também em consideração as 16 covariáveis físico-químicas não-controladas medidas em cada parcela do experimento.

3.0.1 Análise Exploratória: Produção

Codificando as variáveis x_1 = Hibrido = Tratamentos, x_2 = Stand = Número de plantas em unidade experimental e y_1 = Rendimento , traçamos um gráfico de dispersão Rendimento X Stand e box-plots Rendimento X Hibrido conforme Figura 3.1.
Codificação das variáveis físico-químicas:
z_1 = % Argila (m/V), z_2 = Ph Água, z_3 = Índice SMP, z_4 = P (mg/dm3), z_5 = K (mg/dm3), z_6 = % M.O. (m/V), z_7 = Al (cmolc/dm3), z_8 = Ca (cmolc/dm3), z_9 =

Mg (cmolc/dm3), z_{10} = H+Al (cmolc/dm3), z_{11} = CTC (cmolc/dm3), z_{12} = Sat CTC (Bases), z_{13} = Sat CTC (Al), z_{14} = Relação Ca/Mg, z_{15} = Relação Ca/K, z_{16} = Relação Mg/K. Para a análise dos modelos, retiramos aqui as covariáveis z_{14},z_{15} e z_{16}, pois elas representam combinações matemáticas de outras variáveis.

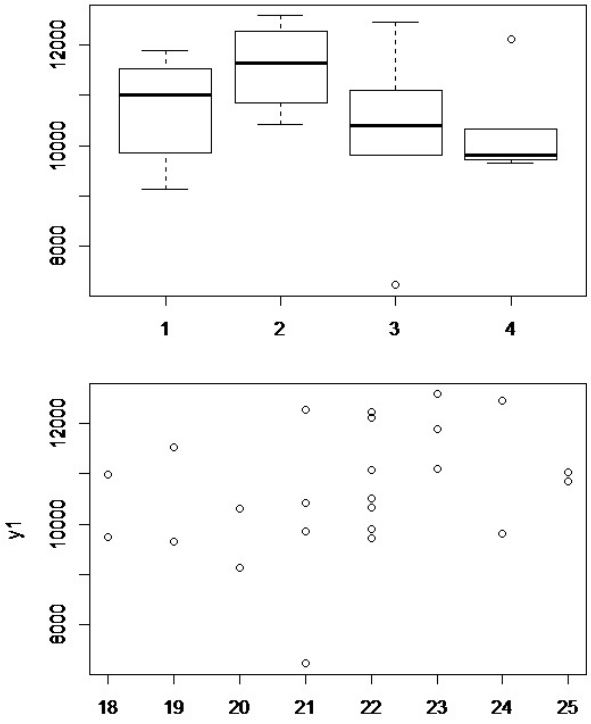

Figura 3.1: Produção x Hibrido e Produção x Stand

O gráfico de dispersão (Fig 3.1) sugere alguma relação, indicando que a variável Stand provavelmente influencia na Produção, podendo esta influência ser significativa ou não. Já os box-plots indicam pequenas diferenças entre os tratamentos (híbridos) na Produção e essa diferença pode ser significativa ou não, dependendo da variabilidade dos modelos.

Para visualisar de forma rápida, a influência ou não das covariáveis do solo na dispersão de Produção em função dos híbridos e Stand, escreveremos a variável produção y_1 num modelo clássico ANOVA em função das variáveis x_1, x_2 e dos blocos; e usaremos

o vetor de resíduos deste modelo para fazer gráficos de dispersão com as covariáveis do solo.

Os gráficos de dispersão descritos nas Figuras 3.2 e 3.3 indicam de forma geral uma moderada influência das covariáveis do solo na dispersão residual do modelo ANOVA. Esta influência pode inflacionar o erro tipo II do modelo induzindo o pesquisador a não detectar diferenças significativas entre tratamentos ou covariáveis, quando na verdade elas existem.

Além disso, detectamos uma leve/moderada correlação linear entre as covariáveis do solo. Esta estrutura de correlação entre elas pode interferir na convergência dos modelos devido a não ortogonalidade destas covariáveis.

3.1 Comparação entre modelos

O modelo ANOVA para a produção de milho $Y_{ij} = b_j + \mu_i + \beta x_2 + \epsilon_{ij}$ com $i = 1,2,3,4$ e $j = 1,2...,6$ onde μ_i são as médias de cada tratamento, b_j são os blocos e ϵ_{ij} os erros aleatórios com distribuição $N(0,\sigma^2)$, apresentou os resultados descritos na tabela 3.1. O teste F não detecta diferenças significativas nos tratamentos e na variável STAND e, escrevendo num modelo linear simples (Tabela 3.2), temos estimativas não significativas do vetor de parâmetros.

Para aplicar a modelagem conjunta da média e dispersão aos dados , suponha $V(\mu) = 1$, função de ligação identidade para o modelo da média e distribuição Gama com função de ligação logarítmica para o modelo da dispersão. Note que não está sendo suposto que o modelo da média é conhecido, pois para usar a modelagem conjunta da média e dispersão necessita-se somente do conhecimento das funções de variância e de ligação.

Observe também que outras funções de variância e de ligação, para o modelo da média, poderiam ter sido usadas. Para as principais distribuições conhecidas, pertencentes

à família exponencial, existem funções de ligação e de variância apropriadas, entretanto podem existir situações em que não se está certo sobre qual distribuição usar, ou seja, não se sabe qual a distribuição correta. Nessas situações podem-se usar as funções de ligação e de variância conhecidas (McCullagh; Nelder, 1989); sendo as melhores aquelas que fornecerem o melhor ajuste aos dados.

No modelo duplo, adotando a estratégia de análise descrita no final do capítulo 2, primeiramente fixamos um modelo maximal para média. Propomos as covariáveis x_1, x_2, $x_1 * x_2$ e b para o modelo maximal da média (código R no apêndice) pois nenhuma interação com as covariáveis do solo foi significativa, em nenhum dos modelos.

Para o modelo de dispersão, como não temos muitos graus de liberdade, decidimos incluir as covariáveis no modelo uma a uma. Adequando as situações convergência do modelo, análise exploratória de dados e significância de parâmetros, encontramos covariáveis candidatas z_7, e z_{10}. Agora, fixado o modelo parcimonioso para média , detectamos uma curvatura no diagnóstico de resíduos e acrescentamos um efeito quadrático na covariável z_{10}.

A descrição dos parâmetros do modelo duplo estão na Tabela 3.3. A modelagem simultânea foi fundamental, pois reduziu a probabilidade do erro tipo II . Basta, por exemplo, calcularmos o coeficiente de variação (CV) de alguns dos parâmetros no modelo linear simples e no modelo duplo. No modelo linear simples , CV(Híbrido 3)=1,15 e CV(Híbrido3*Stand)=1,27 . Já no modelo duplo, CV(Híbrido 3)=0,308 e CV(Híbrido3*Stand)=0,3205 . Além disso, se compararmos o erro padrão entre as duas tabelas, percebemos um erro padrão menor para o modelo MLGD.

Analisando o modelo duplo, o modelo convergiu sendo que no modelo da média detectamos diferença significativa entre híbrido 3 versus híbrido 1, indicando menor produtividade no híbrido 3 abaixo dos híbridos 1, 2 e 4, sendo que não detectamos diferença significativa entre os híbridos 1, 2 e 4. Além disso, a iteração (Hibrido 3)*STAND foi detectada significativa, indicando que o número de plantas na parcela pode interferir

na produtividade do Híbrido 3. Ressalta-se que o modelo ANOVA não detectou diferenças significativas nos tratamentos e o modelo linear simples não encontrou diferenças significativas nos parâmetros.

Já para o modelo de dispersão, a interpretação é que, o aumento de uma unidade da variável z_7 = Alumínio indicam um aumento na dispersão. Por outro lado, o aumento de uma unidade da variávei z_{10} = Hidrogênio + Alumínio indica um efeito quadrático na dispersão, podendo aumentar ou diminuir dependendo da quantidade de H+Al colocada na parcela. Esta análise sugere que o elemento Alumínio ou Alumínio combinado com Hidrogênio tem uma forte influência na variabilidade do experimento

Tabela 3.1: ANOVA: Produção de Milho

FV	GL	SQ	MSQ	F	Valor-p
Blocos	5	11160044	2232009	2,121	0,139
Híbridos	3	7079998	2359999	2,242	0,140
Stand	1	2272810	2272810	2,159	0,170
Hibrido*Stand	3	3648061	1216020	1.155	0.370
Resíduo	11	11578012	1052547		

Tabela 3.2: Modelo Linear Simples: Produção de Milho

Parâmetros	Estimativa	Erro Padrão	Valor-P
Intercepto	7422,81	4586,05	0.13383
Bloco 2	1750,42	738,43	0.03712
Bloco 3	2719,07	842,40	0.00805
Bloco 4	1705,37	853,75	0.07111
Bloco 5	2065,73	1134,46	0.09591
Bloco 6	1753,99	957,03	0.09402
Híbrido 2	11493,76	8475,37	0,20224
Híbrido 3	-6850,16	7882,07	0,40337
Híbrido 4	4443.43	7010,33	0,53914
Stand	78,16	225,39	0,73531
(Hibrido 2)*Stand	-482,10	382,59	0,23371
(Hibrido 3)*Stand	279,62	355,00	0,44753
(Hibrido 4)*Stand	-236,42	334,57	0,49449

A análise de resíduos do modelo duplo (figura 3.4, com os 4 primeiros gráficos

Tabela 3.3: MLGD: Produção de Milho

		Média	
Parâmetros	Estimativa	Erro Padrão	Valor-P
Intercepto	6355,7955	2308,3850	0.018777785
Bloco 2	1564,9358	528,5860	0,012961195
Bloco 3	1964,2345	514,2225	0,002844055
Bloco 4	1662,2998	353,8370	0,000652330
Bloco 5	1125,7158	690,8369	0,131483088
Bloco 6	1722,4354	430,8883	0,002095468
Híbrido 2	5992,7437	5982,7348	0,338027174
Híbrido 3	-25369,6167	7809,8005	0,007756798
Híbrido 4	6214,6264	4059,5047	0,154035780
Stand	151,8243	115,6497	0,215980932
(Hibrido 2)*Stand	-238,0559	271,4340	0,399219234
(Hibrido 3)*Stand	1101,5152	353,0924	0,009756543
(Hibrido 4)*Stand	-336,7280	196,7572	0,115023395
		Dispersão	
Parâmetros	Estimativa	Erro Padrão	Valor-P
Intercepto	36,7200987	4,89032240	$5{,}970482.10^{-14}$
z_7	3,2861955	0,65993244	$6{,}371784.10^{-7}$
z_{10}	-4,4665528	0,94162782	$2{,}101217.10^{-6}$
$I(z_{10}^2)$	0,1626604	0,04333256	$1{,}741956.10^{-4}$

representando os ajustes do Modelo da Média e os 4 últimos gráficos o Modelo da Dispersão) indica uma leve curvatura nos dois modelos (dispersão e média) sugerindo algum efeito de outras covariáveis ou mesmo efeitos quadráticos. Lembrando que os gráficos de ajuste do modelo da média são obtidos baseados na *scaled deviance* de um MLG que tem distribuição assintótica χ^2 e os gráficos de ajuste do modelo da dispersão são baseados na distribuição Gama pra d_i. Como temos um conjunto de dados limitado com poucas observações e altamente sensível quanto à convergência, decidimos por hora ignorar estes pequenos problemas, deixando para artigos futuros com ferramentas mais sofisticadas de análise.

Nelder e Lee (1991) afirmam que a modelagem conjunta da média e dispersão é geral e suficiente para ajustar os modelos de Taguchi. Usando MLG não é preciso usar transformação para os dados. Modelos com resposta contínua, ou na forma de contagem

e proporção, podem ser ajustados usando o mesmo algoritmo. Além disso, o critério de separação pode ser satisfeito pela especificação correta da função de variância no MLG; e parcimônia pode ser encontrada escolhendo funções apropriadas de ligação e covariáveis para os parâmetros dos modelos da média e da dispersão, respectivamente.

Um outro aspecto muito importante da modelagem conjunta da média e dispersão é que esta abordagem permite encontrar, além dos fatores que afetam a média, aqueles que afetam a dispersão. Dessa forma, pode-se escolher valores das covariáveis de modo que a resposta para o modelo da dispersão seja mínima.

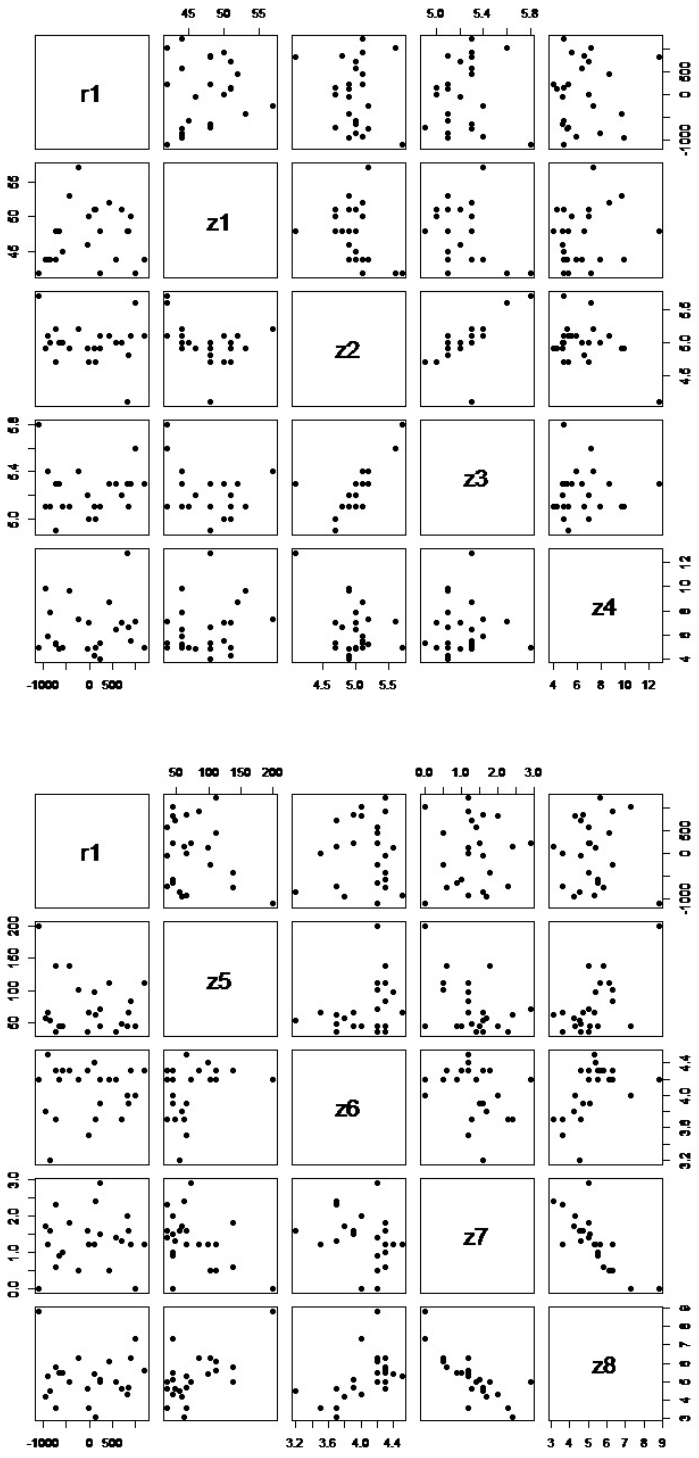

Figura 3.2: Resíduos r_1 versus covariáveis do solo

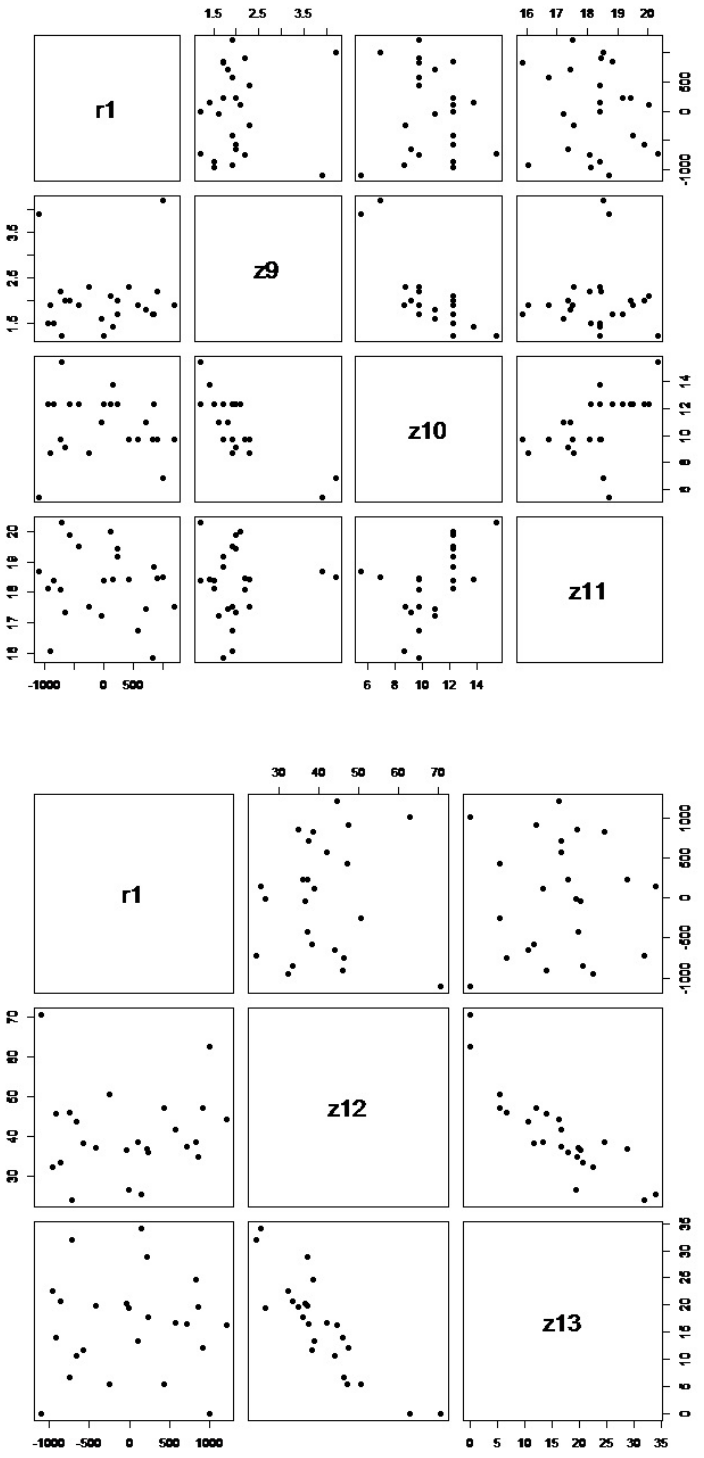

Figura 3.3: Resíduos r_1 versus covariáveis do solo

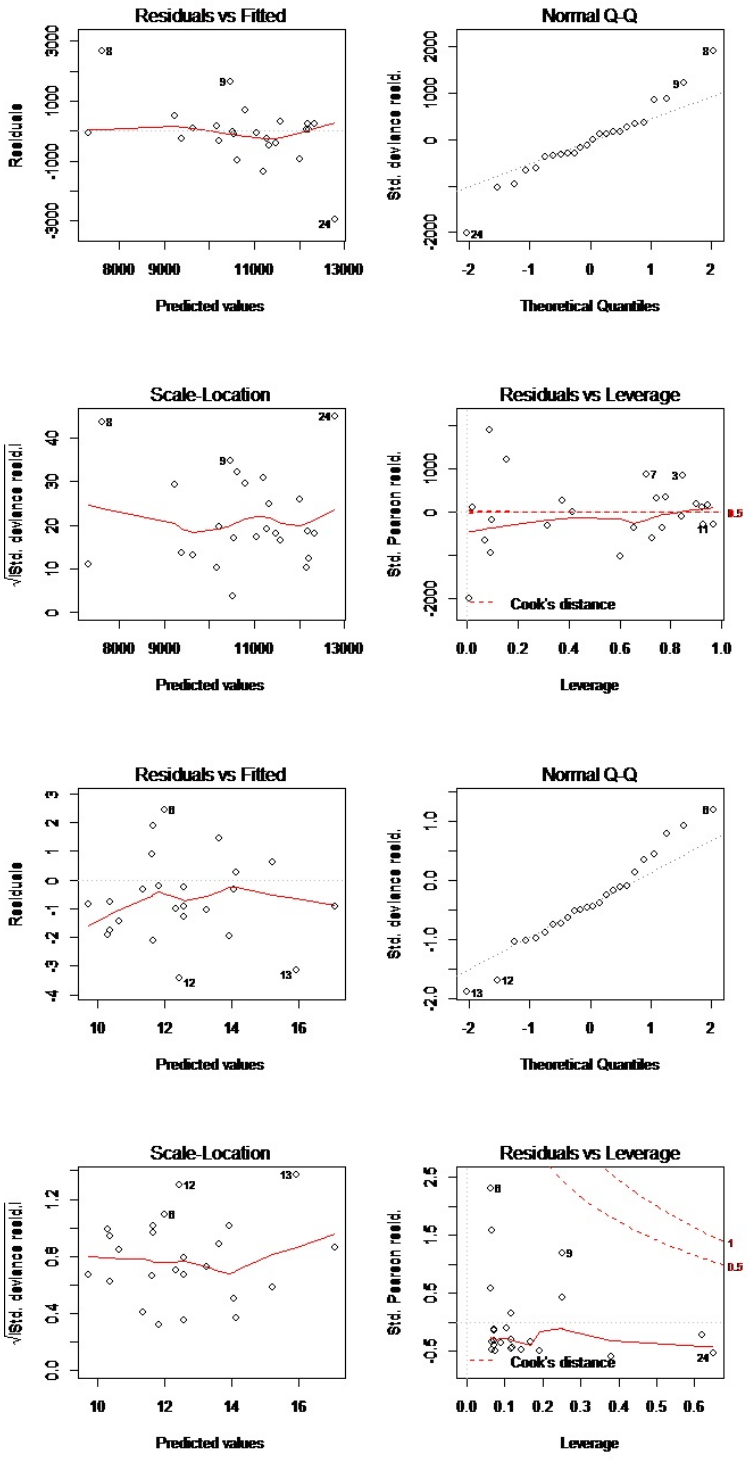

Figura 3.4: Qualidade de ajuste: Modelos da Média e Dispersão

Capítulo 4

Considerações Finais

É cada vez mais recorrente, dentre as pesquisas que utilizam a análise clássica do Planejamento Robusto de experimentos em blocos casualisados com parcelas subdivididas em tratamentos, a ocorrência de várias fontes de variação não-controláveis gerando uma possível sobredispersão nos modelos simples ANOVA. Esta abordagem inflaciona o erro tipo II influenciando o pesquisador a não detectar diferenças significativas para covariáveis ou entre tratamentos, sendo que na verdade elas existem.

A modelagem simultânea de média e dispersão neste caso é altamente relevante pois, além de detectar diferenças significativas entre variáveis ou tratamentos, indica ao pesquisador possíveis caminhos no controle local de covariáveis que influenciam na variabilidade do experimento. Além disso, características fundamentais do Planejamento Robusto são preservadas (BOX (1988)) auxiliando no controle da variabilidade de um experimento mantendo a média de uma característica de qualidade em um valor nominal.

Temos ainda perspectivas, no experimento analisado nesta monografia, para trabalhos futuros, envolvendo as variável Numero de Espigas, que é um processo de contagem, demandando um modelo Poisson duplo. Acrescentamos que temos um número grande de covariáveis de ruído neste experimento, induzindo-nos a testar a aborda-

gem de utilização de Análise de Componentes Principais, com o objetivo de reduzir a dimensionalidade das variáveis de ruído e melhorar o controle da variabilidade.

Referências Bibliográficas

AITKIN, M. (1987). *Modelling variance heterogeneity in normal regression using GLIM. Applied Statistics.* London, v. 36, n. 3, p.332-339.

BOX, G. (1988). *Signal-to-noise ratios, performance criteria and transformations.* Technometrics, Alexandria, v. 30, n. 1, p.1-40.

COCHRAN, W. G. e COX, G. M. (1957). *Experimental designs.* 2.ed. New York: John Wiley Sons.

DEMÉTRIO, C. G. B. (2002). *Modelos lineares generalizados em experimentação agronômica.* Piracicaba: Departamento de Ciências Exatas, ESALQ/USP.

HINDE, J. e DEMÉTRIO, C. G. B. (1998). *Overdispersion: models and estimation. Computational Statistics Data Analisys.* Amsterdam, v. 27, n. 2, p. 151-170.

LEE, Y. e NELDER, J. A. (1998). *Generalized linear models for athe analysis of quality-improvement experiments.* The Canadian Journal of Statistics, Ottawa, v. 26, n. 1, p. 95-105.

LEE, Y., NELDER, J. A., e PAWITAN, Y. (2006). *Generalized linear models with random effects:Unified analysis via H-likelihood.* London: Chapman Hall.

McCULLAGH, P. e NELDER, J. A. (1989). *Generalized linear models.* 2nd.ed. London: Chapman Hall.

NAIR, V. N. (1992). *Taguchi's parameter design: a panel discussion.* Technometrics,34(2), 127-161.

NELDER, J. A. e LEE, Y. (1991). *Generalized linear models for the analysis of Taguchi-type experiments. Applied Stochastic Models and Data Analysis.* Chichester, v. 7, p. 107-120.

NELDER, J. A. e PREGIBON, D. (1987). *An extended quasi-likelihood function.* Biometrika,74, 221-232.

NELDER, J. A. e WEDDERBURN, W. M. (1972). *Generalised linear models.* Journal of the Royal Statistical Society Series A, London, v. 135, n. 3, p. 370-384.

PINTO, E. R. e PONCE DE LEON, A. C. (2006). *Modelagem conjunta da média e dispersão de Nelder e Lee como alternativa aos métodos de Taguchi.* Pesquisa Operacional, v.26, n.2, p.203-224.

PREGIBON, D. (1984). *Review of generalized linear models by P. McCullagh and J. Nelder.* The Annals of Statistics.

SHOEMAKER, A. C., TSUI, K., e LEÓN, R. V. (1988). *Signal-to-noise ratios, performance criteria, and transformation: discussion.* Technometrics, Alexandria, v. 30, n. 1, p. 19-21.

SMYTH, G. K. e VERBYLA, A. P. (1999). *Adjusted likelihood methods for modelling dispersion in generalized linear models.* Environmetrics, v. 10, p. 696-709.

SMYTH, K. (1989). *Generalized linear models with varing dispersion.* Journal of the Royal Statistical Society Series B, London, v. 51, n. 1, p. 47-60.

TAGUCHI, G. (1985). *Communication in Statistics: Theory and Methods.* Hamilton, v. 14, p. 2785-2801.

VIEIRA, A., LEANDRO, R., DEMETRIO, C., e MOLENBERGHS, G. (2010). *Double generalized linear model for tissue culture proportion data: a Bayesian perspective.* Journal of Applied Statistics, p. 1-15.

VIEIRA, A. M. C. e DEMÉTRIO, C. G. B. (2008). *Modelagem simultânea de média e dispersão e aplicações na pesquisa agronômica.* Piracicaba: Departamento de Ciências Exatas, ESALQ/USP.

VIEIRA, F. M. C., SILVA, I. J. O., BARBOSA-FILHO, J. A. D., VIEIRA, A. M. C., e BROOM, D. M. (2011). *Preslaughter mortality of broilers in relation to lairage and seasons of the year in the subtropical climate.* Poultry Science , v. 90, p. 2127-2133.

WEDDERBURN, R. W. M. (1974). *Quasi-likelihood functions, generalized linear models, and Gauss-Newton method.* Biometrika, London, v. 61, n. 3, p. 439-447.

Apêndice A

Script no R

```
    setwd(choose.dir("C:/Estatistica/Estagio"))
mono <- read.csv("ExperimentoEmBlocosComAnaliseQuimicaDaParcela.csv", sep =
";")
y1<-mono$RENDIMENTO
y2<-mono$ESPIGAS
b<-mono$BLOCOS
x1<-as.factor(mono$HIBRIDO)
x2<-mono$STAND
z1<-mono$Argila
z2<-mono$pH
z3<-mono$Indice
z4<-mono$Pmg
z5<-mono$KMg
z6<-mono$MO
z7<-mono$Al
z8<-mono$Ca
z9<-mono$Mg
```

```
z10<-mono$HAL
z11<-mono$CTC
z12<-mono$SatCTC1
z13<-mono$SatCTC2
Análise Exploratória de Dados
op<-par(mfrow=c(2,1),mar=c(3, 7, 1, 5))
boxplot(y1 x1)
plot(y1 x2)
M1<-aov(y1 x1*x2+b)
summary(M1)
r1<-resid(M1)
pairs(r1 z1+z2+z3+z4,pch=19)
pairs(r1 z5+z6+z7+z8,pch=19)
pairs(r1 z9+z10+z11,pch=19)
pairs(r1 z12+z13,pch=19)
Modelo Linear Simples
M2<-lm(y1 b+x1*x2)
summary(M2)
Calculando os modelos maximais
Média
library(dglm)
mod1<-dglm(y1 b+x1*x2+x1*z1+x2*z1, 1,family=gaussian)
mod2<-dglm(y1 b+x1*x2+x1*z2+x2*z2, 1,family=gaussian)
mod3<-dglm(y1 b+x1*x2+x1*z3+x2*z3, 1,family=gaussian)
mod4<-dglm(y1 b+x1*x2+x1*z4+x2*z4, 1,family=gaussian)
mod5<-dglm(y1 b+x1*x2+x1*z5+x2*z5, 1,family=gaussian)
mod6<-dglm(y1 b+x1*x2+x1*z6+x2*z6, 1,family=gaussian)
```

```
mod7<-dglm(y1 b+x1*x2+x1*z7+x2*z7, 1,family=gaussian)
mod8<-dglm(y1 b+x1*x2+x1*z8+x2*z8, 1,family=gaussian)
mod9<-dglm(y1 b+x1*x2+x1*z9+x2*z9, 1,family=gaussian)
mod10<-dglm(y1 b+x1*x2+x1*z10+x2*z10, 1,family=gaussian)
mod11<-dglm(y1 b+x1*x2+x1*z11+x2*z11, 1,family=gaussian)
mod12<-dglm(y1 b+x1*x2+x1*z12+x2*z12, 1,family=gaussian)
mod13<-dglm(y1 b+x1*x2+x1*z13+x2*z13, 1,family=gaussian)
summary(mod1)
summary(mod2)
summary(mod3)
summary(mod4)
summary(mod5)
summary(mod6)
summary(mod7)
summary(mod8)
summary(mod9)
summary(mod10)
summary(mod11)
summary(mod12)
summary(mod13)
media<-dglm(y1 b+x1*x2, 1,family=gaussian)
summary(media)
Dispersão
disp1<-dglm(y1 b+x1*x2, z1,family=gaussian)
disp2<-dglm(y1 b+x1*x2, z2,family=gaussian)
disp3<-dglm(y1 b+x1*x2, z3,family=gaussian)
disp4<-dglm(y1 b+x1*x2, z4,family=gaussian)
```

```
disp5<-dglm(y1 b+x1*x2, z5,family=gaussian)
disp6<-dglm(y1 b+x1*x2, z6,family=gaussian)
disp7<-dglm(y1 b+x1*x2, z7,family=gaussian)
disp8<-dglm(y1 b+x1*x2, z8,family=gaussian)
disp9<-dglm(y1 b+x1*x2, z9,family=gaussian)
disp10<-dglm(y1 b+x1*x2, z10,family=gaussian)
disp11<-dglm(y1 b+x1*x2, z11,family=gaussian)
disp12<-dglm(y1 b+x1*x2, z12,family=gaussian)
disp13<-dglm(y1 b+x1*x2, z13,family=gaussian)
summary(disp1)
summary(disp2)
summary(disp3)
summary(disp4)
summary(disp5)
summary(disp6)
summary(disp7)
summary(disp8)
summary(disp9)
summary(disp10)
summary(disp11)
summary(disp12)
summary(disp13)
Modelo Final
modelofinal1<-dglm(y1 b+x1*x2, z7+z10,family=gaussian,maxit=500)
summary(modelofinal1)
modelofinal<-dglm(y1 b+x1*x2, z7+z10+I(z10^2),family = gaussian)
summary(modelofinal)
```

```
Ajuste
par(mfrow = c(2,2))
plot(modelofinal$dispersion.fit)
par(mfrow = c(2,2))
plot(modelofinal)
```