Comparação de duas abordagens utilizando modelos mistos para um experimento de cana-de-açúcar

Renata Alcarde[1], Simone Daniela Sartorio[1], Emídio Cantídio Almeida de Oliveira[2], Sônia Maria De Stefano Piedade[1], César Gonçalves de Lima[3], Clarice Garcia Borges Demétrio[1], Paulo Cesar Ocheuze Trivelin[4]

1 Introdução

Muitos são os campos de estudo onde se efetuam várias observações sobre a mesma unidade experimental ao longo do tempo. Planejamentos do tipo longitudinais são frequentemente utilizados em diversas áreas de pesquisa, porque permitem avaliar mudanças globais ou individuais ao longo do tempo (SINGER; NOBRE; ROCHA, 2009). Como as medidas são repetidas de modo sistemático, nas mesma unidades experimentai, espera-se que exista uma correlação não nula entre as medidas e uma heterocedasticidade das variâncias nas diversas ocasiões.

Uma abordagem que vem sendo muito utilizada para análise de dados longitudinais é a modelagem com modelos mistos. Segundo Pinheiro e Bates (2000), um modelo com ambos efeitos fixos e aleatórios é chamado de modelo de efeitos mistos, ou simplesmente modelo misto.

Ao escolher um modelo de regressão para descrever o comportamento de uma variável resposta de acordo com variáveis de controle, tem-se a opção de usar modelos polinomiais, que são lineares nos parâmetros. Aumentando-se a ordem desses polinômios, pode-se obter aproximações cada vez mais precisas da verdadeira função de regressão, geralmente, não-lineares. Modelos não lineares de efeitos mistos são uma estensão dos modelos lineares de efeitos mistos, e permitem que a função de regressão não linear dependa de efeitos fixos e aleatórios.

[1]LCE - ESALQ/USP; Contatos: alcarde83@yahoo.com.br, sisartorio@yahoo.com.br
[2]LSO, ESALQ/USP
[3]Departamento de Ciências Básicas - FZEA/USP
[4]CENA/USP.

Neste trabalho, utilizaram-se duas abordagens mistas, linear e não-linear, para modelar os dados de produção de matéria seca de cana-de-açúcar (kg/ha), provenientes de um experimento longitudinal, casualizado em blocos.

2 Materiais e Métodos

Os dados utilizados neste trabalho foram obtidos junto ao departamento de Solos e Nutrição de Plantas da Escola Superior de Agricultura "Luiz de Queiroz", de um experimento instalado seguindo o delineamento casualizado em blocos (4 blocos) com quatro tratamentos (doses de nitrogênio - 0, 40, 80 e 120) aplicados na cana-planta em que cada parcela foi observada em cinco ocasiões.

O modelo proposto na literatura para dados com tais características é originalmente não linear, geralmente um modelo logístico, que possui parâmetros de fácil interpretação e carac-terísticas já conhecidas, de tal forma a fornecer informações sobre a resposta de interesse acumulada, por meio da primeira derivada do modelo ajustado. Neste trabalho, utilizaram-se duas abordagens para modelar o conjunto de dados: a primeira abordagem envolve um modelo linear misto e a segunda, um modelo não linear misto.

O modelo linear misto pode ser escrito como (LAIRD; WARE, 1982 e MARSHALL; BARÓN, 2000):

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \quad \text{para} \quad i = 1, 2, \ldots, N$$

em que $\mathbf{y}_i$ representa um vetor $(n_i \times 1)$ de respostas da i-ésima unidade experimental ou indivíduo, $\mathbf{X}_i$ é uma matriz $(n_i \times p)$ de especificação dos efeitos fixos (conhecida), $\boldsymbol{\beta}$ é vetor $(p \times 1)$ de parâmetros (efeitos fixos), $\mathbf{Z}_i$ é uma matriz $(n_i \times q)$ de especificação (conhecida) dos efeitos aleatórios, $\mathbf{b}_i$ é um vetor $(q \times 1)$ de efeitos aleatórios, tal que $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{G})$ e $\boldsymbol{\varepsilon}_i$ é um vetor $(n_i \times 1)$ de erros aleatórios em que $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \mathbf{R}_i)$, logo $Var(\mathbf{Y}_i) = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i$.

As matrizes de especificação $\mathbf{X}_i$ e $\mathbf{Z}_i$ podem ser diferentes e variar entre unidades experimentais, estendendo o modelo para o caso de dados não balanceados em relação ao tempo. As matrizes $\mathbf{Z}_i$ podem conter quaisquer covariáveis que afetem diferentemente as unidades experimentais, e a forma de especificação das matrizes $\mathbf{X}_i$ é bastante similar àquela utilizada nos modelos de regressão.

Segundo Pinheiro e Bates (2000), a mais comum aplicação dos modelos não lineares mistos é para dados com medidas repetidas, em particular, dados longitudinais. O modelo não-linear de efeitos mistos para medidas repetidas, considerando o i-ésimo indivíduo, pode ser representado como:

$$\mathbf{y}_i = f(\mathbf{X}_i, \boldsymbol{\beta}) + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i$$

com a mesma estrutura do modelo linear misto descrito anteriormente e f é uma função não-linear em pelo menos um dos componentes do vetor de parâmetros. Um modelo mais geral, para o qual o efeito aleatório é parte do componente não linear do modelo pode ser representado como:

$$\mathbf{y}_i = f(\mathbf{X}_i, \boldsymbol{\beta}, \mathbf{Z}_i, \mathbf{b}_i) + \boldsymbol{\varepsilon}_i$$

Para o modelo logístico:

$$f(\mathbf{X}_i, \boldsymbol{\beta}, \mathbf{Z}_i, \mathbf{b}_i) = \frac{\phi_{1i}}{1 + \exp\left(-\frac{(t_{ij} - \phi_{2i})}{\phi_{3i}}\right)}$$

em que $\boldsymbol{\phi}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i$, tal que $\boldsymbol{\phi}_{1i}$ representa a assíntota do modelo, $\boldsymbol{\phi}_{2i}$ representa o tempo até o indivíduo atingir metade da resposta assintótica e $\boldsymbol{\phi}_{3i}$ representa, aproximadamente, 3/4 da resposta assintótica (parâmetro de escala).

Os parâmetros dos modelos não-lineares, geralmente, têm uma interpretação física natural e mesmo quando derivados empiricamente, incorporam conhecimento, características teóricas dos dados, tais como assíntotas e monotonicidade. Um modelo não-linear, geralmente, usa menos parâmetros do que um modelo linear, e fornecem previsões mais confiáveis para a variável resposta fora do intervalo observado.

Considerando as duas abordagens, inicialmente tomou-se o modelo maximal para a parte fixa, estimando-se os componentes de variância pelo método da máxima verossimilhança restrita (MMVR) e buscando uma estrutura variâncias e covariâncias adequada. Tendo selecionado o modelo, pelo teste da razão de verossimilhanças e pelos Critérios de Informação de Akaike (AIC) e Bayesiano (BIC), tal modelo foi reestimado utilizando o método da máxima verossimilhança e a significância dos termos fixos do modelo foi testada utilizando-se a estatística de Wald.

As análises foram feitas utilizando o pacote *nlme: Linear and Nonlinear Mixed Effects Models* para o programa R.

3 Resultados e discussão

Os gráficos de perfis apresentados na Figura 1 estão relacionados ao ajuste do modelo linear misto considerando um efeito aleatório do intercepto para bloco e tratamento, um coeficiente linear e um coeficiente quadrático para cada tratamento.

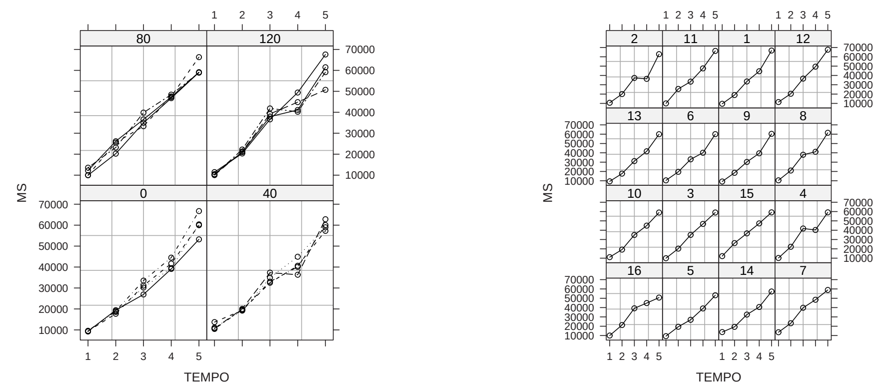

Figura 1: Gráfico de perfis individuais e por tratamento de produção de matéria seca (kg/ha)

Estruturas de variâncias foram testadas, e o modelo final inclui um componente aleatório somente no intercepto e matriz $\mathbf{R}_i$ autoregressiva de ordem 1 com heterogeneidade de variâncias. O teste de Wald indicou a necessidade de uma curva polinomial para cada tratamento.

O modelo logístico inicial considerou para cada parâmetro um componente fixo para efeitos de tratamento e bloco, e um componente aleatório para captar a variabilidade entre os indivíduos submetidos a cada um dos tratamentos. O modelo final selecionado contém apenas um componente aleatório na assíntota e considerando a parte fixa desse modelo, ajustaram-se curvas distintas para cada tratamento.

Os conjuntos de gráficos apresentados nas Figuras 2 e 3 mostram aleatoridade nos resíduos e ausência de *outliers*, indicando bons ajustes dos modelos linear e não linear, respectivamente, além das curvas obtidas de cada modelo por tratamento.

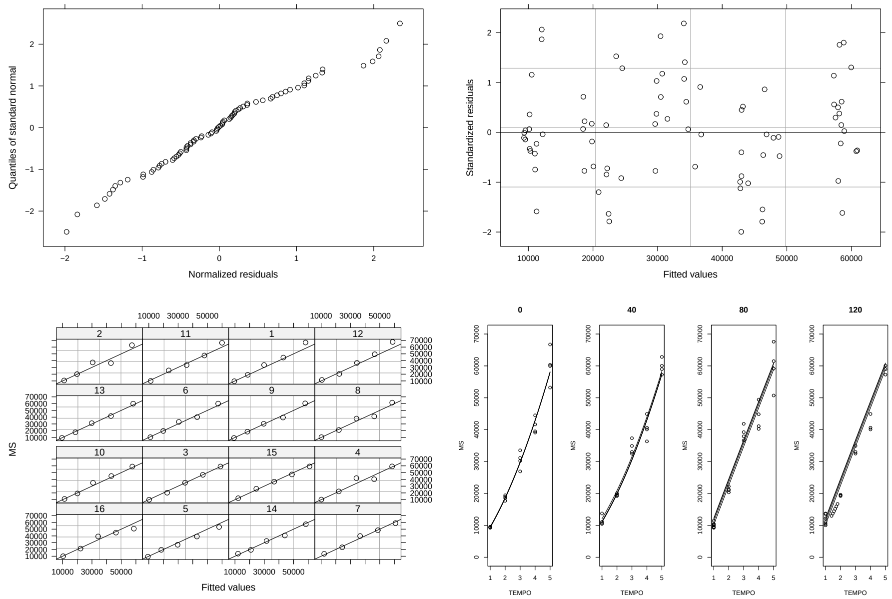

Figura 2: Gráficos: quantil-quantil para os resíduos, valores ajustados *versus* resíduos estudentizados, valores ajustados *versus* valores observados e curvas ajustadas para cada tratamento considerando um modelo linear misto.

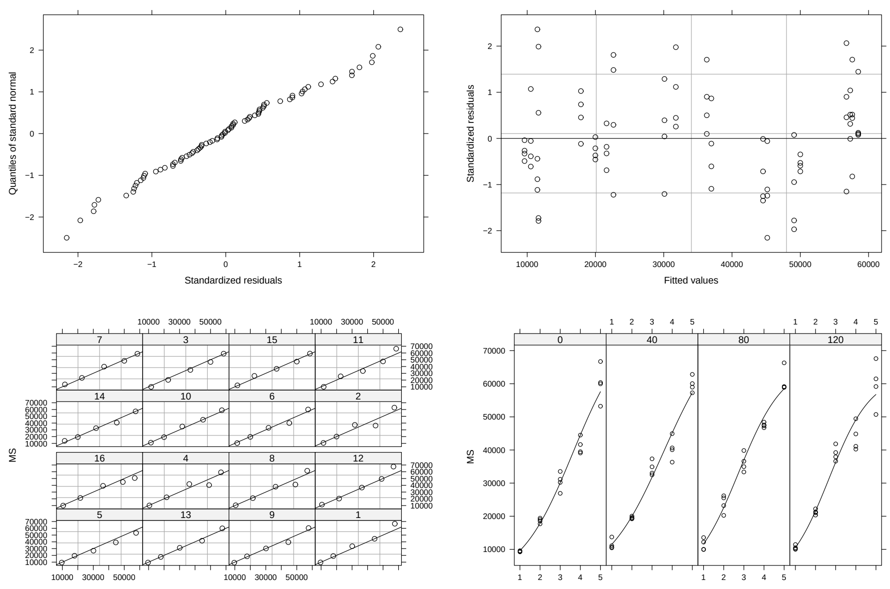

Figura 3: Gráficos: quantil-quantil para os resíduos, valores ajustados *versus* resíduos estudentizados, valores ajustados *versus* valores observados e curvas ajustadas para cada tratamento considerando o modelo logístico misto.

A partir das informações apresentadas na Tabela 1 é possível comparar as duas abordagens. Pela estatística BIC o modelo escolhido foi o não linear, quando utilizado o MMV para a estimação dos parâmetros, pois esse critério penaliza os modelos com menor número de parâmetros. Entretanto, considerando o MMVR o modelo selecionado foi o linear. Embora o modelo linear tenha apresentado menor AIC, o modelo não linear possui menor número de parâmetros e melhor interpretação do fenômeno.

Tabela 1: Informações sobre os ajustes dos modelos mistos linear e não linear realizados pelos métodos de máxima verossimilhança (MMV) e máxima verossimilhança restrita (MMVR)

Modelo	n^o parâmetros	MMV		MMVR	
		AIC	BIC	AIC	BIC
Linear	20	1511,158	1563,563	1305,463	1353,300
Não Linear	17	1511,335	1547,066	1466,853	1500,146

4 Conclusão

A quantidade de produção de matéria seca de cana-de-açúcar ao longo do tempo pode ser modelada utilizando-se regressão linear ou não linear, ambas com termos aleatórios. Embora a modelagem linear seja mais simples e tem apresentado um menor AIC, ele precisa de um número maior de parâmetros que o do modelo não linear. Este último, por sua vez, mostrou-se mais atraente, pois as estimativas dos seus parâmetros têm uma melhor interpretação, uma vez que derivada de primeira ordem é amplamente utilizada para a verificação do acúmulo de matéria seca no período de interesse.

Bibliografia

LAIRD, N.M.; WARE, J.H. Random effects models for longitudinal data. **Biometricks**, Washington, v. 38, p. 963-974, 1982.

MARSHALL, G.; BARÓN, A.E. Linear discriminant models for unbalanced longitudinal data. v.19, p.1969-1981, 2000.

PINHEIRO, J.C.; BATES, D.M. Mixed-Effects Models in S and S-PLUS. New York: Springer-Verlag, 528p., 2000.

SINGER, J.; NOBRE, J.S.; ROCHA, F.M.M. Análise de dados longitudinais. Apostila, 158p., 2009.