Revista da Estatística UFOP, Vol III(3), 2014, ISSN 2237-8111
Edição especial: 59ª Reunião Anual da Regional Brasileira da Sociedade Internacional de Biometria - RBRAS
Trabalhos apresentados na versão poster.

Utilização de modelos marginais na análise de dados longitudinais irregulares em relação ao tempo

César Gonçalves de Lima[1]
Michele Barbosa[2]
Valdo Rodrigues Herling[3]

1. Introdução

Dados longitudinais surgem de experimentos com medidas repetidas, em que as medidas são feitas nas mesmas unidades experimentais ao longo do tempo. Geralmente esses dados exibem alguma correlação serial e heterocedasticidade de variâncias e na sua análise são utilizadas, por exemplo, a análise multivariada de perfis, a análise baseada na especificação de modelos lineares marginais e de modelos lineares (ou não lineares) mistos.

Quando o intervalo entre duas medidas consecutivas não é constante ao longo do estudo, os dados são chamados *irregulares* em relação ao tempo e as análises uni e multivariada de perfis, por exemplo, tornam-se inadequadas.

O presente trabalho procura evidenciar as vantagens da análise de um conjunto de dados longitudinais irregulares, baseadas na especificação de um modelo linear marginal para dados correlacionados.

2. Material e métodos

Os dados utilizados neste trabalho resultaram de um experimento de pastejo instalado no campus da USP, Pirassununga-SP. O objetivo da pesquisa foi obter informações concretas da presença de animais em pastejo sobre o capim Mombaça (*Panicum maximum* Jacq), quando submetido a dois períodos de descanso (PDD = 28 e 35 dias) e três pressões de pastejo (PP = 3,3; 4,1 e 4,9%), avaliando a recuperação da área de pastejo em quatro piquetes e em cinco ocasiões (1, 2, 3, 5 e 7). Das diversas variáveis relativas à produção e à qualidade avaliadas durante o experimento, foi escolhida a altura (cm) média das touceiras do piquete. Os dados resultantes são classificados como irregulares em relação ao tempo porque algumas parcelas foram avaliadas aos 28, 56, 84, 140 e 196 dias e outras, aos 35, 70, 105, 175 e 245 dias.

[1] Departamento de Ciências Básicas, FZEA/USP, Pirassununga/SP - E-mail: cegdlima@usp.br
[2] Instituto de Ciências Sociais Aplicadas. UNIFAL–MG
[3] Departamento de Zootecnia, FZEA/USP, Pirassununga/SP.

Na análise dos dados utilizou-se o modelo linear marginal para dados correlacionados que pode ser escrito (Vonesh, 2013) como:

$$\boldsymbol{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \quad i = 1,\ldots, n \text{ unidades experimentais} \tag{1}$$

em que $\boldsymbol{y}_i$ é um vetor $p \times 1$ de respostas avaliadas em p ocasiões; $\mathbf{X}_i$ é uma matriz $p \times s$ de delineamento; $\boldsymbol{\beta}$ é um vetor $s \times 1$ de parâmetros desconhecidos; $\boldsymbol{\varepsilon}_i$ é um vetor $p \times 1$ de erros aleatórios independentes (entre unidades experimentais) e identicamente distribuídos com $\boldsymbol{\varepsilon}_i \sim N_p\big(\mathbf{0}, \boldsymbol{\Sigma}_i(\boldsymbol{\theta})\big)$ e $\boldsymbol{\theta}$ é um vetor $d \times 1$ de parâmetros de variâncias e covariâncias que define a estrutura de variabilidade e correlação entre as medidas feitas em cada unidade experimental. Assumindo que

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_n \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix}, \ \boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{bmatrix} \text{ e } \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{\Sigma}_1(\boldsymbol{\theta}) & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2(\boldsymbol{\theta}) & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\Sigma}_n(\boldsymbol{\theta}) \end{bmatrix}$$

pode-se escrever o modelo (1) de forma sucinta como:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \text{ com } \boldsymbol{\varepsilon} \sim N_N\big(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})\big) \tag{2}$$

em que $\boldsymbol{y}$ e $\boldsymbol{\varepsilon}$ são vetores $N\times1$; $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ é uma matriz $N\times N$ positiva definida e $N = \sum_{i=1}^{n} p_i$ representa o número total de observações.

Este modelo é bastante geral e oferece diversas opções para analisar dados correlacionados, como as técnicas familiares de: análise de variância multivariada, análise de variância para medidas repetidas, análise de curvas de crescimento etc. Dependendo da abordagem, a matriz $\boldsymbol{\Sigma}_i(\boldsymbol{\theta})$ pode assumir diferentes estruturas, como: completamente parametrizada ou não estruturada (UN), simetria composta (CS), autorregressiva de 1ª ordem (AR(1)) etc.

A estimação do modelo pode ser feita usando o método dos mínimos quadrados generalizados (MQG), das equações de estimação generalizadas (EEG), da máxima verossimilhança (MV) ou da máxima verossimilhança restrita (MVR). No presente estudo foram utilizados os métodos MV e MVR.

A busca pelo melhor modelo iniciou-se com a especificação de um modelo completo para $E(\boldsymbol{y})$**,** formado por retas distintas para cada combinação dos níveis dos fatores PP e PDD e por uma estrutura completamente parametrizada para $\boldsymbol{\Sigma}_i(\boldsymbol{\theta})$. Mantendo o mesmo modelo para $E(\boldsymbol{y})$**,** outras estruturas para $\boldsymbol{\Sigma}_i(\boldsymbol{\theta})$ foram testadas. No ajuste dos modelos utilizou-se o método MVR e nas comparações entre eles, o teste da razão de verossimilhanças (TRV), pois os modelos são encaixados.

Após a escolha da melhor estrutura de $\boldsymbol{\Sigma}_i(\boldsymbol{\theta})$, buscou-se o melhor modelo para $E(\boldsymbol{y})$ realizando comparações entre os parâmetros em $\boldsymbol{\beta}$, utilizando teste de Wald e testes da razão

de verossimilhanças. No ajuste desses modelos utilizou-se o método MV e na comparação de modelos não encaixados foram utilizados os critérios de informação de Akaike (AIC) e o bayesiano de Schwarz (BIC). Em todas as análises, inclusive na análise de diagnóstico, utilizou-se o *proc mixed* do SAS (Littel et al., 2006) que disponibiliza diversas estruturas para $\boldsymbol{\Sigma}_i(\boldsymbol{\theta})$.

3. Resultados e discussões

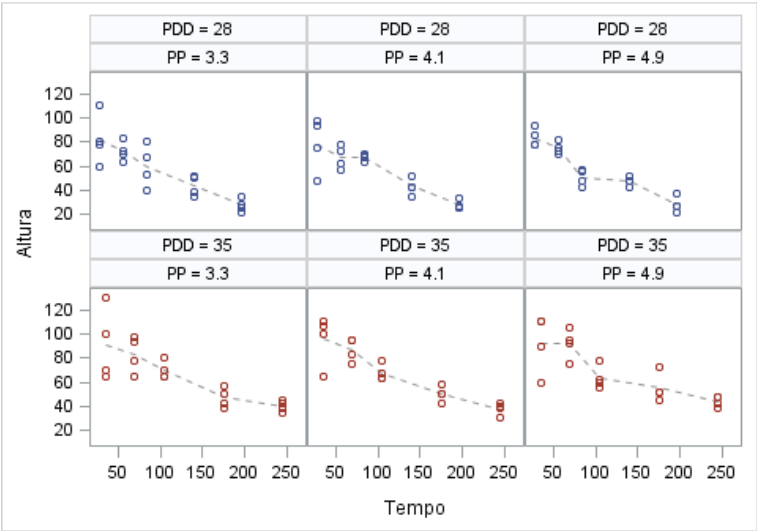
O uso do modelo (2) na análise dos dados é justificado pela presença de heterogeneidade das variâncias nas diversas ocasiões (Tabela 1), bem como pela presença de correlações altas e distintas.

Tabela 1. Variâncias (diagonal), covariâncias (acima da diagonal) e correlações amostrais (abaixo da diagonal), por período de descanso (PDD)

	PDD = 28 dias					PDD = 35 dias				
Ocasião	1	2	3	5	7	1	2	3	5	7
1	1705,03	1178,67	667,01	96,92	-581,66	2998,96	2385,15	1143,00	458,65	-86,00
2	0,96	881,82	484,65	74,98	-417,52	0,95	2106,25	1042,85	393,50	-85,65
3	0,81	0,82	400,74	25,98	-245,43	0,82	0,89	647,44	185,18	-56,80
5	0,36	0,39	0,20	41,81	-17,85	0,72	0,73	0,62	136,18	1,78
7	-0,89	-0,89	-0,77	-0,17	250,81	-0,29	-0,34	-0,41	0,03	29,71

O uso de retas distintas para explicar os comportamentos das alturas médias em função do tempo (dias) de todas as combinações PDD×PP e a presença de heterogeneidade de variâncias ao longo do tempo é sugerido pela análise da Figura 1.

Figura 1. Gráfico de dispersão das alturas médias das touceiras em função do tempo (dias), por nível de PDD e PP.



Dentre as estruturas de variâncias e covariâncias avaliadas, a estrutura CSH, que admite variâncias diferentes nas diversas ocasiões e correlações iguais entre tempos, foi a escolhida (Tabela 2). É uma estrutura mais simples que a UN, que é utilizada nas análises multivariadas de perfis; porém é mais complexa que a VC, que é utilizada nas análises de regressão linear simples. A necessidade de usar matrizes de covariâncias distintas para dados de parcelas submetidas a diferentes níveis de PDD foi inicialmente descartada (p-valor > 0,10).

Tabela 2. Estatísticas de ajuste para escolha de estrutura da matriz $\boldsymbol{\Sigma}_i(\boldsymbol{\theta})$

Número	Estrutura	-2logV[(1)]	AIC	BIC	#par[(2)]	Compara	TRV[(3)]	g.l.[(4)]	p-valor
1	UN	872.5	902.5	920.1	15				
2	TOEPH	884.4	902.4	913.0	9	1x2	11.9	6	0.0642
3	CSH	888.1	900.1	907.2	6	2x3	3.7	3	0.2957
4	TOEP	923.0	933.0	938.9	5	3x4	34.9	1	< 0.0001
5	CS	929.1	933.1	935.5	2	3x5	41.0	4	< 0.0001
6	VC	934.6	936.6	937.8	1	3x6	46.5	5	< 0.0001

[(1)] –2log da verossimilhança do modelo
[(2)] número de parâmetros da estrutura de covariâncias
[(2)] estatística do teste da razão de verossimilhanças
[(4)] número de graus de liberdade associado a TRV

Testes de Wald evidenciaram somente o efeito significativo dos níveis de PDD nos interceptos e coeficientes angulares das retas, indicando que o melhor modelo envolva somente duas retas distintas para os diferentes níveis do fator PDD.

As retas ajustadas que explicam o comportamento da altura média das touceiras em função dos diferentes níveis de período de descanso (28 e 35 dias, respectivamente) podem ser escritas como:

$$\hat{y}_l = 87{,}6737 - 0{,}3063t_l, \text{ para } t_l = 28, 56, 84, 140 \text{ e } 196 \text{ dias.}$$
$$\hat{y}_l = 98{,}9578 - 0{,}2437t_l, \text{ para } t_l = 35, 70, 105, 175 \text{ e } 245 \text{ dias.}$$

Testes de Wald indicam diferenças (p-valor < 0,05) entre os interceptos e as inclinações das duas retas, permitindo concluir que: as touceiras dos piquetes submetidos a períodos de descanso de 35 dias tinham uma altura média superior no início do experimento e tiveram uma taxa de diminuição da altura média ao longo do tempo menor que o das touceiras dos piquetes submetidos a períodos de descanso de 28 dias.

Um exame dos gráficos apresentados na Figura 2 evidencia a ausência de pontos atípicos e a adequação da distribuição normal aos erros, indicando uma boa qualidade do ajuste do modelo final.

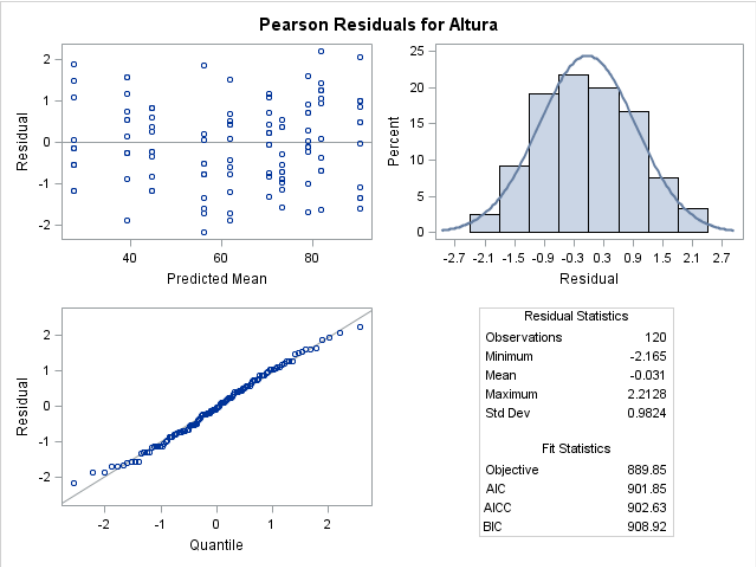


Figura 2. Diagnóstico do modelo final

4. Conclusões

O uso de modelos marginais para dados longitudinais irregulares em relação ao tempo é bastante eficiente, porque facilita a busca por uma estrutura adequada e parcimoniosa, que explica bem a heterogeneidade de variâncias e a correlação serial, que são características comuns desse tipo de dados.

5. Referências

LITTELL, R. C.; MILLIKEN, G. A.; STROUP, W. W.; WOLFINGER, R. D.; SCHABENBERGER, O.. **SAS® for Mixed Models**. 2nd Edition. Cary, NC: SAS Institute Inc. 2006.

VERBEKE, G., MOLENBERGHS, G. **Linear Mixed Models for Longitudinal Data**. New York, Springer. 2000.

VONESH, E. F. **Generalized Linear and Nonlinear Models for Correlated Data: Theory and Applications Using SAS®**. Cary, NC: SAS Institute Inc. 2012.