https://eventos.utfpr.edu.br//sicite/sicite2018

Uso de modelos mistos lineares para análises de experimentos fatoriais usando o software R

Use of linear mixed models for analysis of factorial experiments using the software R

Naldisya Drosdrocky Gonçalves
naldisyadrosdrocky@gmail.com
Universidade Tecnológica Federal do Paraná, Londrina, Paraná, Brasil
Joelmir André Borssoi
jborssoi@utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Londrina, Paraná, Brasil
Gabriel Lourenço Rodrigues
gabriel_lrodrigues@hotmail.com
Universidade Tecnológica Federal do Paraná, Londrina, Paraná, Brasil

RESUMO

Abrangendo uma gama de problemas dos mais diversos, os modelos mistos estão disseminados em várias áreas, da matemática à modificação de genes nas culturas agrícolas. A análise multivariada através de modelos mistos, tem grande aplicação devido, principalmente, a sua grande flexibilidade de acomodação das intraunidades amostrais, comumente encontradas em dados longitudinais. A análise multivariada se refere a todas as técnicas estatísticas que, simultaneamente, analisam múltiplas medidas sobre indivíduos ou objetos sob investigação. Este trabalho buscou realizar uma revisão bibliográfica a fim de se compreender o funcionamento das técnicas de análises multivariadas, a partir de modelos de regressão, tanto os lineares clássicos quanto os de efeitos mistos. Além disso, comparar os ajustes de modelos de efeitos mistos usando diferentes distribuições de probabilidade, tanto para o efeito aleatório, quanto para o erro aleatório. Para isso foi utilizado o pacote *heavy,* disponibilizado no *Software R*. Para comparar os ajustes, foram aplicados os critérios de informação de Akaike (AIC), Bayesiano (BIC) e Log-Verossimilhança.

PALAVRAS-CHAVE: Modelos mistos. Técnicas multivariadas. *Software R*.

ABSTRACT

Covering a range of problems from the most diverse, mixed models are widespread in many areas, from mathematics to gene modification in agricultural crops. The multivariate analysis through mixed models, has great application due mainly to its great flexibility of accommodation of the sample intraunidades, commonly found in longitudinal data. Multivariate analysis refers to all statistical techniques that simultaneously analyze multiple measures on individuals or objects under investigation. This work aimed to perform a bibliographical review in order to understand the operation of the multivariate analysis techniques, from regression models, both classical and mixed effects linear. In addition, we compare the mixed effects model settings using different probability distributions, both for the random effect and for the random error. In order to compare the adjustments, we used the Akaike (AIC), Bayesian (BIC) and Log-likelihood information criteria.

KEYWORDS: Mixed models. Multivariate techniques. *Software R*.

Recebido: 31 ago. 2018.
Aprovado: 04 out. 2018.

1 INTRODUÇÃO

Cotidianamente, no meio acadêmico são realizados experimentos e testes de hipótese, que fazem o uso de diversas variáveis dentro de uma mesma amostra de indivíduos. Em muitos casos, essas variáveis são medidas em múltiplos períodos de tempo e quando isso acontece, pode ser chamado de estudo longitudinal. Dentre as técnicas que modelam bem as diferentes variáveis, estão os modelos mistos lineares (MML).

Em algumas situações, quando medidas são tomadas no tempo, pode-se fazer o uso da análise em parcelas subdivididas no tempo, mas na maioria das vezes isto não será a melhor opção para realizar análises de dados com medidas repetidas. Este modelo pressupõe que o erro da parcela, que engloba o fator de tratamentos e o erro da sub-parcela, onde estão postos os tempos e a interação: tempos versus tratamento, tenham distribuição normal, sejam independentes e identicamente distribuídos.

Ainda, segundo Xavier (2000), a condição válida para o uso de parcelas subdivididas, com medidas no tempo, é que a matriz de covariância seja de simetria composta. Essa simetria sugere que a variável aleatória esteja igualmente correlacionada e tenha a mesma variância, considerando medidas em diferentes instantes. Respeitar essa adoção é importante pois caso a matriz de covariância não apresente a forma simétrica composta, o teste F será não exato e ocorrera uma inflação sobre o efeito da subparcela, erro que segundo Pinheiro (1994), poderia ser sanado com o uso de modelos mistos.

Segundo Costa (2003), quando a variável resposta é observada ao longo do tempo (dados longitudinais), pode haver uma correlação entre as observações e isso deverá ser levado em conta na estimação de parâmetros. Para tanto, é interessante a inserção de um “efeito” que estime as variabilidades não consideradas no modelo e que podem influenciar nos resultados.

Ela necessitará de um bom embasamento teórico, pois sem estar conceitualmente firme sobre as especificidades e individualidades de restrições ao qual sua técnica é baseada não se pode extrair os dados corretos ou a amplitude máxima alcançada pelo método.

O modelo linear de regressão mais clássico é expresso por: $y = X\beta + \epsilon$, em que y representa o vetor de dimensões $nx1$, de dados observados, X, de dimensões nxp, é a matriz de delineamento, β , de dimensões $px1$, é um vetor de parâmetros desconhecidos de efeitos fixos e ϵ é o vetor de dimensão $nx1$, de erros aleatórios. Segundo Laird e Ware (1982) o modelo misto linear normal para respostas contínuas assume a seguinte forma:

$$y_i = X_i\beta + Z_i b_i + \in_i, \qquad (1)$$

$$b_i \overset{iid}{} N_q(0, D) \text{ e } \in_i \sim N_{m_i}(0, R_i), \quad i = 1, \ldots, n \qquad (2)$$

em que y_i representa o vetor aleatório m_i-dimensional das respostas observadas para o i-nésimo indivíduo ou grupo, X_i e Z_i são matrizes de planejamentos $(m_i \times p)$ e $(m_i \times q)$, respectivamente, β é um vetor p-dimensional

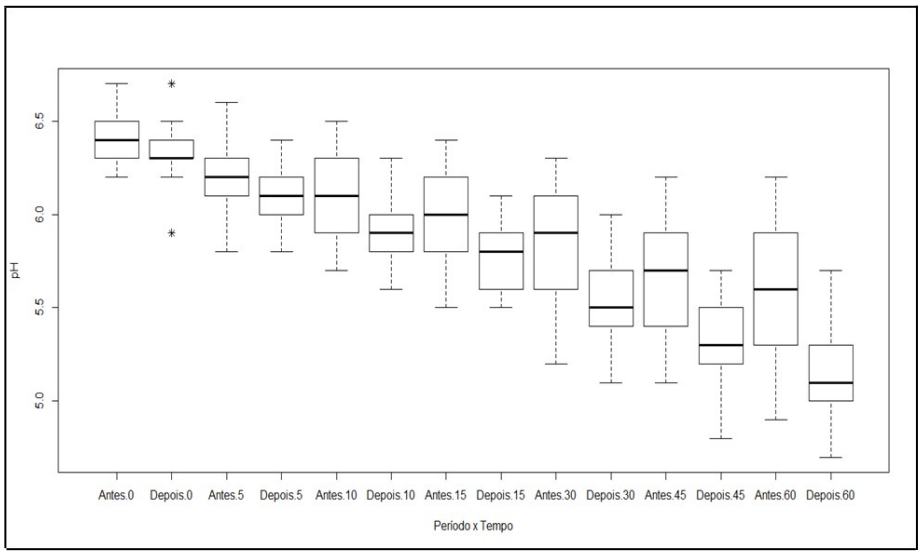
de efeitos fixos, b_i denota um vetor q-dimensional de efeitos aleatórios e $\in_i$ representa um vetor de erros.

2 RESULTADOS

Para ilustrar a aplicação da modelagem por meio de modelos mistos foram utilizados dados provenientes de um estudo realizado na Faculdade de Odontologia da Universidade de São Paulo para avaliar o efeito do uso contínuo de uma solução para bochecho no pH da placa bacteriana dentária. Com essa finalidade, o pH da placa dentária retirada de 21 voluntários antes e depois do uso de uma solução para bochecho foi avaliado ao longo de 60 minutos após a adição de sacarose ao meio em que as unidades experimentais foram colocadas. Os dados foram obtidos por Grande et al. (1998).

A seguir são apresentados gráficos box plot para analisar descritivamente a variação de pH, de acordo com o período e o tempo após a utilização da solução para bochecho. Observando a Figura 2, pode-se perceber que há maior homogeneidade no período “depois”, onde a diferença de amplitude dos quartis é menor que em relação ao período “antes”, isso nos mostra que os valores de pH, diferem menos na placa bacteriana, quando o indivíduo já efetuou o

Figura 1: Gráfico Blox Plot, para os períodos de antes e depois do uso da solução.



bochecho. Outra indicação é o fato dos valores de pH, diminuírem gradativa e diretamente quanto ao uso da solução.

No software R (R CORE TEAM, 2017), além da análise descritiva, e com o auxílio do pacote heavy, mais precisamente utilizando a função heavy.lme, foram ajustados vários modelos com diferentes configurações, partindo de um modelo completo, que continha todas as variáveis e interação entre elas, conforme a equação (2):

$$y = \overbrace{(\beta ¿¿ 0 + \beta_1 Período + \beta_2 Tempo + \beta_3 periodo * Tempo)}^{ef.fixos} + \overbrace{b_o + b_1 Período + b_1 Tempo}^{ef.aleatórios} \quad (3)$$

Partindo do modelo completo, as combinações de ajuste dos modelos seguintes foram feitas retirando-se a(s) variável(is) não significativa(s) (p-valor maior que o nível de significância de 5%).

Nos ajustes foram consideradas as distribuições normal e t de Student (com diferentes graus e liberdade (df)), tanto para o efeito aleatório quanto para o erro aleatório.

Como critério de seleção de modelos foram escolhidos: BIC, AIC e Log-Verossimilhança. Os parâmetros β,foram estimados pela função de verossimilhança, conforme a Tabela 2. Para os critérios de AIC e BIC, quanto menor é o seu valor, melhor é o ajuste. Já para Log-Verossimilhança, quanto maior o valor, melhor será o ajuste.

Tabela 1: Valores dos critérios escolha dos modelos e suas verossimilhanças.

Modelo	AIC	BIC	Log-Verossimilhança
Normal[1]	17.38945	20.66788	-3.389451
Normal [2]	-70.87222	-68.06214	82.87222
Normal[3]	-67.96031	-65.61858	77.96031
Student[4] (df=3)	8.666449	11.95316	5.333551
Student[5] (df=4)	8.662846	11.94128	5.337154
Student[6] (df=10)	8.674729	11.95316	5.325271

Considerando a Tabela 1, percebe-se que o modelo Normal², foi o mais bem-sucedido, alcançando os menores valores de BIC e AIC e os maiores valores de Log-Verossimilhança. Ele considerou a interação "Período*Tempo", não significativa.

O segundo melhor modelo é o Normal³, onde apenas variável "Tempo" foi considerada significante, em seguida encontramos o modelo Normal¹, que apresentou os piores valores da distribuição normal, sendo superior apenas aos valores de t-Student.

Na distribuição de t de Student, foram considerados diferentes graus de liberdade, para se testar como o modelo se comportava em relação as restrições de variação. No segundo modelo, com df=4, a interação Período*Tempo, mostrou-se significativa, já o terceiro modelo (com df=10) devido a menor restrição de variação, se aproxima da normal, apresentando as interações, "Tempo" e "Período", como significativas e o ajuste torna-se pior (longe do nível de significância). Por último analisou-se o primeiro modelo (com df=3), onde apenas a variável período foi considerada significativa, o que em níveis práticos o torna semelhante ao modelo Student (df=10).

1 Normal 1: Modelo completo, com as variáveis Período*Tempo.

2 Normal 2: Modelo com as variáveis Período + Tempo.

3 Normal 3: Modelo com a variável Tempo.

4 Student (df=4): Modelo com quatro graus de liberdade.

5 Student (df=10): Modelo com dez graus de liberdade.

6 Student (df=3): Modelo com três graus de liberdade.

3 CONCLUSÃO

Com o desenvolvimento deste trabalho, foi possível compreender o uso de modelos de regressão para representar as variações de um fenômeno (representado por uma variável), em função de outras variáveis e que o processo de modelagem (ajuste de modelos) pode ser feito de diversas formas, envolvendo diferentes tipos de modelos, de acordo com a natureza dos dados. Neste caso, os modelos de efeitos mistos.

Observando os resultados apresentados, nota-se a capacidade que os modelos mistos têm de analisar experimentos fatoriais, englobando os efeitos aleatórios, permitindo testar a significância dos mesmos para propor um modelo final.

Pôde-se perceber, a partir da revisão bibliográfica, que há muitos estudos passíveis da aplicações de modelos mistos na área de Engenharia Ambiental, porém, muitas vezes esta abordagem não é utilizada, deixando de levar em consideração a possível correlação entre as medidas (ou observações) feitas no tempo. Isto pode mascarar resultados e levar o pesquisador a conclusões errôneas.

Também é importante destacar o uso do *software* R, como ferramenta para análises estatísticas, visto que, além de proporcionar uma enorme variedade de análises, é gratuito.

3.1 PERSPECTIVAS FUTURAS

Como a proposta inicial do plano de trabalho é para 12 meses e este prazo não terminou, pretende-se ainda melhorar e ampliar as análises estatísticas, além de fazer aplicações em dados relacionados à Engenharia Ambiental.

REFERÊNCIAS

HAIR JR., J.F.; WILLIAM, B.; BABIN, B.; ANDERSON, R.E. Análise multivariada de dados. 6.ed. Porto Alegre:Bookman, 2009.

LEITE, Mayana Silva Bessa et al. Comparação entre metodologias de amostragem de água para quantificação de variáveis limnológicas em ambiente lótico. Ambiente e Água, Itapetinga, v. 12, n. 1, p.136-145, 16 maio 2016. Mensal.COMPANHIA DE TECNOLOGIA DE SANEAMENTO AMBIENTAL DO ESTADO DE SÃO PAULO- CETESB. Guia nacional de coleta e preservação de amostras: água, sedimento, comunidades aquáticas e efluentes líquidos. São Paulo: CETESB; Brasília: ANA, 2011. 325 p.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. NBR 9897: Planejamento de amostragem de efluentes líquidos e corpos receptores : Procedimento. Rio de Janeiro, 1987. 14 p.

COSTA, Silvano Cesar da. Modelos Lineares Generalizados Mistos para dados longitudinais. 2003. 107 f. Tese (Doutorado) - Curso de Agronomia, Universidade de São Paulo, Piracicaba, 2003.

DEMIDENKO. E. Mixed models: theory and applications with R. 2.ed. Wiley & Sons. 2013. 717p.

WU, L. Mixed Effects Models for Complex Data. Monografhs on Statistics and Applied Probability 113. A Chapman & Hall Book. (2010).

BORSSOI, Joelmir André. Modelos Mistos Lineares elípticos com erros de medição. 2014. 143 f. Tese (Doutorado) - Curso de Estatística, Universidade de São Paulo, São Paulo, 2014.

XAVIER, L. H. Modelos univariados e multivariados para análise de medidas repetidas e verificação da acurácia do modelo uni variado por meio de simulação. Dissertação (mestrado em estatística e experimentação agronômica), Escola superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2000.

LAIRD, N. M., and WARE, J. H. Random-effects models for longitudinal data. Biometrics 38, 4 (1982), 963–974.

NASCIMENTO, Alan Rener Borges. Análise da produtividade de cultivares de café utilizando modelos lineares mistos. 2016. 48 f. Tese (Doutorado) - Curso de Bacharel, Departamento de Estatística, Universidade de Brasília, Brasília, 2016.

BATISTA, João Luiz F.. Verossimilhança e Máxima Verossimilhança. Centro de Métodos Quantitativos, Departamento de Ciências Florestais, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Campus Piracicaba.2009.

RAMOS, Pedro Luiz; ACHCAR, Jorge Alberto; RAMOS, Eduardo. Método Eficiente Para Calcular Os Estimadores De Máxima Verossimilhança Da Distribuição Gama generalizada. Revista Brasileira de Biometria, São Paulo, v. 32, n. 2, p.267-281, fev. 2014.

SEARLE, S. R., CASELLA, G., and MCCULLOCH, C. E. Variance components, vol. 391. John Wiley & Sons, 2009.

GRANDE et al. Efeitos do uso contínuo de solução para bochecho sobre o pH e o conteúdo mineral da placa bacteriana. Revista da Pós-graduação da Faculdade de Odontologia da Universidade de São Paulo, São Paulo, v. 2, n. 0, p.143-147, nov. 1998.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

PINHEIRO, J. C. Topics in mixed effects models. Tese (doutorado em estatística), University of Wisconsin, Madison, 1994.

AGRADECIMENTOS

Agradeço a Deus e ao meu orientador, Joelmir André Borssoi, bem como meu colega de trabalho, Gabriel Lourenço. À Universidade Tecnológica Federal do Paraná (UTFPR-Campus Londrina) e ao CNPQ, pela oportunidade de desenvolver esta experiência.