USO DE *MULTIPLE IMPUTATION* EM DADOS DE BOVINOS DE CORTE

ALFREDO RIBEIRO DE FREITAS[1], LUCIANA M. GUARDIA[2]

[1] Bolsista do CNPq
[2] Bacharel de Estatística da UFSCar

RESUMO – valores perdidos de dados de pesos de bovinos de corte foram estimados por meio da estratégia de múltipla *imputation*. De cada animal foram analisadas nove pesagens, gerando um arquivo completo. A partir deste foi gerado um arquivo incompleto (AI), com os dados submetidos à perda do tipo monótono. Para substituir os dados perdidos no AI foram comparados o método da Regressão Paramétrica (RP), que usa as variáveis prévias como covariáveis e o Não-Paramétrico de scores de "Propensão" (NP), que agrupa as observações com base nestes escores. O método RP foi o mais adequado para estimar valores perdidos submetidos à perda do tipo monótono.

PALAVRAS-CHAVE: bovinos de corte, dados completos, dados incompletos, medidas repetidas, procedimento MI

USE OF MULTIPLE IMPUTATION IN REPEATED MEASURE OF BEEF CATTLE

ABSTRACT- Missing values of cattle weight were estimated by multiple imputation. From each animal were analyzed nine weighing originating a complete data set. In this data set was imposed a missing data pattern of monotone type generating an incomplete data set. In order to replace missing value in the incomplete data set, two methods were compared: Parametric Regression (PR), that uses the previous variables as covariates and the Non-Parametric Regression propensity scores (NP). The PR method was the most adequate for estimating missing values of cattle weight when the monotone type for missing data is considered.

KEYWORDS: beef cattle, complete data, incomplete data, MI procedure, repeated measures

INTRODUÇÃO

Dados de pesos de bovinos obtidos ao longo da vida do animal é de fundamental importância na produção animal para obter estimativas de parâmetros genéticos, ajustar curvas de crescimento, análises de medidas repetidas, etc. Para se utilizar eficientemente essas análises deve-se atentar para algumas características inerentes a dados de crescimento de bovinos como medidas repetidas: a) as pesagens são irregulares no tempo; b) possuem estrutura incompleta; c) as avaliações adjacentes são mais estreitamente correlacionadas do que as demais; d) a resposta dos animals em função do tempo tem variância crescente.

A ocorrência de estrutura incompleta em dados de crescimento de bovinos resulta em perda de informações e prejuízos em diversas análises. A maioria dos procedimentos do SAS, por exemplo, exclui das análises todas as observações do animal que possui dados perdidos. No melhoramento genético animal, os modelos de regressão aleatórias, de grande interesse para estimar covariâncias com a idade, são bastante sensíveis a problemas amostrais e de perdas de dados (Nobre et al. 2003). Para solucionar tais problemas, está em desenvolvimento a técnica de "Imputação" Múltipla (Rubin, 1996), a qual substitui cada valor perdido pela média de um conjunto de valores plausíveis representando a incerteza sobre o valor correto. Uma vez que considera todos os dados existentes como informações à *priori*, o valor estimado representa o valor mais plausível de acontecer.

O objetivo deste trabalho foi comparar dois métodos de "imputação" múltipla quanto à eficiência em estimar dados de pesos de bovinos Nelore, submetidos à perdas do tipo monótono: Regressão Paramétrica (RP), que usa as variáveis prévias como covariáveis e o Não-Paramétrico de escores de "Propensão" (NP), que agrupa as observações com base nestes escores.

MATERIAL E MÉTODOS

Foi utilizado dados de nove pesagens: ao nascimento (P_O) e oito pesagens posteriores (P1 a P8), realizadas em intervalos trimestrais, até os dois anos de idade de uma amostra de 25 animais Nelore, oriundos da ABCZ. A partir desta amostra, foi organizado um arquivo incompleto, impondo aos

dados perdas do tipo monótono (para a sequência de pesos P_0, P_1, ..., P_8, avaliado em um animal, se um peso P_j é perdido, implica que os pesos subseqüentes P_k, $k > j$, são todos perdidos).
Para substituir os dados perdidos no arquivo incompleto, foi utilizado o procedimento MI (Rubin,1996) por meio do SAS (SAS, 2000) utilizando-se Regressão Paramétrica (RP), que usa as observações não perdidas do animal como covariáveis e o método Não-Paramétrico de Escores de Propensão (NP), em que valor perdido é substituído por um conjunto de valores plausíveis que representa a incerteza acerca do valor verdadeiro a ser fornecido (imputado).

Para medir a eficiência dos métodos de "imputação" múltipla em estimar dados submetidos à perdas do tipo monótono, a similaridade dos dados dos dois conjuntos (completo e incompleto) foi avaliados por três critérios por meio de procedimentos do SAS (SAS, 2000): a) intervalos de confiança; b) estrutura de (co)variância da variabilidade das medidas dentro dos animais; foram testadas as seguintes estruturas: CS, AR(1), ARMA(1,1), CSH, FA1(1), HF e UN (Wolfinger, 1993). Para testar essas estruturas foram usados os seguintes critérios fornecidos pelo procedimento MIXED do SAS (SAS, 2000)): distribuição de χ^2, AIC (Akaike's Information Criterion), SBC (Schwarz's Bayesian Criterion) e *-2LOGR* e c) ajuste de curvas de crescimento pelo modelo de Von Bertallanfy.

RESULTADOS E DISCUSSÃO

Os intervalos de confiança (Tabela 1) associados às medias dos pesos indicam que ambos os métodos: Regressão Paramétrica (RP) e Propensão (NP), foram similares; contudo, o NP produziu para o peso P7 um intervalo excessivamente grande (106,14 = 253,70 - 359,84). As médias dos valores obtidos pelos dois métodos não diferiram muito da média observada, implicando que o conjunto de valores plausíveis gerados para substituir um valor perdido, representaram adequadamente a incerteza do valor correto. A eficiência desses métodos em substituir valores perdidos em dados longitudinais foram comprovados em dados clínicos envolvendo dados longitudinais (Cook, 1996; Mazumdar, 1999).

Das estruturas de covariâncias avaliadas: CS, AR(1), ARMA(1,1), CSH, FA1(1), HF e UN, a Não Estruturada (UN), foi a mais adequada para representar a variabilidade das medidas dentro dos animais, tanto no arquivo completo quanto no incompleto. A Tabela 2 apresenta os valores dos critérios de ajuste proporcionados pelo MIXED do SAS para a covariância UN. O resultado χ^2_{44} = 521,04 < 0,0001 para os dados completos, por exemplo, mostra que o modelo que estima a matriz UN é significativamente (P < 0,0001) melhor do que o modelo simples nulo. Por outro lado, a similaridade dos valores para -2LOGR, AIC E SBC, entre os arquivos completo e o incompleto com os valores "imputados" pelo método de Regressão Paramétrica (RP), mostra que as duas estruturas de covariâncias UN são praticamente as mesmas, indicando que o método RP estimou adequadamente os dados.

A Figura 1 apresenta os limites de confiança, superior e inferior, com, 95% de probabilidade, obtidos do modelo de Von Bertalanffy para os dados de pesos do nascimento até 720 dias de idade. Verifica-se uma sobreposição dos intervalos de confiança para ambos os arquivos: completo (linha cheia) e incompleto (linha pontilhada) para o modelo de Regressão Paramétrica (esquerda). Contudo, no método de Propensão (direita), os intervalos de confiança diferem nas idades superiores a 450 dias. Este fato, comprova as observações já efetuadas nas Tabelas 1 e 2, ou seja, o método RP foi o mais adequado para estimar valores perdidos submetidos à perda do tipo monótono a exemplo.

CONCLUSÕES

O método de Regressão Paramétrica estimou adequadamente dados perdidos de pesos de bovinos analisados como medidas repetidas e submetidos à perda do tipo monótono.

REFERÊNCIAS BIBLIOGRÁFICAS

COOK, N. R. Accounting for missing data in clinical trials of blood pressure (BP) and hypertension. **Controlled Clinical Trials**. v17, n.2, p.595-605, Supplement 1, April 1996.

MAZUMDAR, S., LIU,K.S., HOUCK, P. R. et al. Intent-to-treat analysis for longitudinal clinica. ltrials: coping with the challenge of missing values. **Journal of Psychiatric Research**, v.33, n.2, p.87-95,1999.

NOBRE,P.R.C. MISZTAL,I.TSURUTA,S. et al. Genetic evaluation of growth curves in Nellore cattle by multiple-trait and random regression models. **Journal of Animal Science**, v.81,p. 927-932,2003.

RUBIN, D. B. Multiple Imputation after 18+ Years. **Journal of the American Statistical Association**, v.91,p.473-489.1996.

SAS Institute Inc. Statistical Analysis System user's guide. Version 8.2 ed. Cary: SAS Institute, USA, 2000.

WOLFINGER, R. **Covariance structure selection in general mixed models**. Community of Statistics - Simulation, v.22, n.4, p.1079-1106, 1993.

TABELA 1. Média dos pesos observados (MO); médias e limite inferior (LI) superior (LS) com 95% de probabilidade obtidos dos métodos de Regressão Paramétrica e de Propensão

		Regressão Paramétrica			Propensão		
Pesos	M0	Média	LI	LS	Média	LI	LS
P1	61,48	60,83	50,14	71,51	60,67	48,88	72,46
P2	112,64	114,54	103,32	125,76	114,64	102,79	126,48
P3	154,00	157,09	141,32	172,86	156,60	141,90	171,29
P4	194,92	196,00	174,71	217,29	193,68	172,56	214,79
P5	240,28	242,11	213,99	270,23	241,38	214,77	267,99
P6	272,28	274,64	248,75	300,54	273,24	249,54	296,93
P7	304,20	301,74	277,07	326,42	306,77	253,70	359,84
P8	347,20	342,78	310,02	375,54	367,81	338,48	397,15

TABELA 2. Critérios de ajuste proporcionados pelo procedimento MIXED do SAS para a estrutura de covariância Não Estruturada obtida dos dados completos (AC) e incompletos (AI) usando os métodos de Regressão Paramétrica e de Propensão

Arquivo	***-2LOG R***	***AIC***	**SBC**	χ^2
AC: completo	580,3	670,3	725,1	χ^2_{44} = 521,04 < 0,0001
AI: Regressão Paramétrica	537,5	627,5	682,3	χ^2_{44} = 565,30 < 0,0001
AI: Propensão	1.745,1	1.835,1	1.890,0	χ^2_{44} = 531,31 < 0,0001

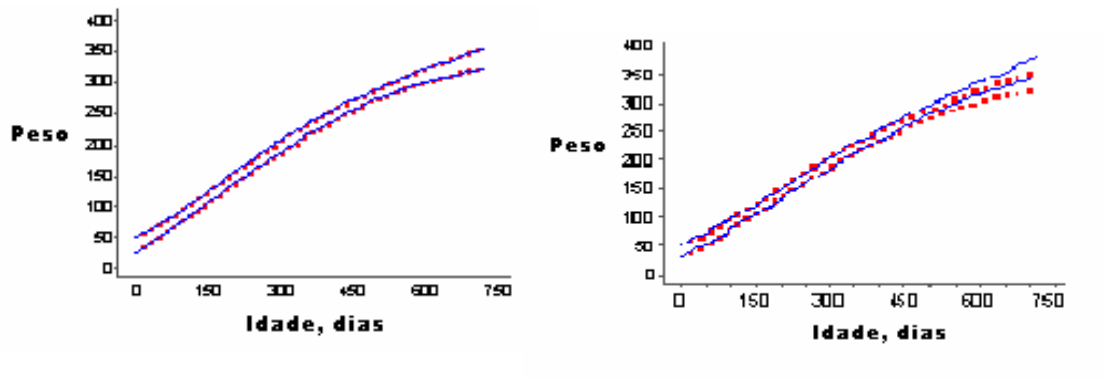

FIGURA 1. Limites de confiança, superior e inferior, com, 95% de probabilidade, obtidos do modelo de Von Bertalanffy para dados de pesos do nascimento até 720 dias de idade de Nelore. A linha cheia e a pontilhada, indicam, respectivamente, o ajuste para os dados de pesos do arquivo completo e incompleto para o método de Regressão Paramétrica (esquerda) e Propensão (direita)