

CONTRAPRODUCÊNCIA E ERROS NA CONVERSÃO DE VARIÁVEIS CONTÍNUAS EM CATEGÓRICAS PARA A ANÁLISE DE VARIÂNCIA

Carlos Alberto Oliveira de Matos (carlos@itapeva.unesp.br) - Campus Experimental de Itapeva/Unesp

RESUMO: A análise de variância simples (ANOVA) para fatores categóricos nada mais é do que um caso especial da ANOVA com fatores contínuos que é geralmente denominada de regressão. Simulações com métodos Monte Carlo (MMC) foram efetuadas com o objetivo de mostrar que a conversão de variáveis contínuas em categorias (procedimento comum nos ensaios com madeira) é quase sempre contra-produtiva e frequentemente, indutora de erro. De um experimento, onde a porcentagem de folhas num aglomerado de folhas e partículas de bambu foi convertida em cinco categorias equidistantes entre 0% e 100%, foram obtidos os dados de tração paralela para cada categoria. O modelo linear obtido na ANOVA com fatores contínuos foi simulado em oito diferentes condições de variabilidade (coeficientes de variação categóricos entre 1% e 25%). Os resultados indicam que a abordagem da conversão de variáveis contínuas em categóricas, embora atraente, pode induzir conclusões equivocadas: os mesmos dados que na ANOVA com fatores contínuos produziram modelos lineares onde o coeficiente de determinação (R^2) era baixo e portanto inadequado para descrever eficientemente a variação observada, produziram uma porcentagem de comparações múltiplas significativas ($\alpha = 0,05$) elevada na ANOVA com fatores categóricos. Por exemplo, na condição de maior variabilidade, os mesmos dados que determinaram médias de coeficientes de determinação inferiores a 0,40 produziram em média 40% de comparações múltiplas significativas.

Palavras-chave: simulação, métodos Monte Carlo (MMC), tração paralela, ANOVA

CONVERSION OF CONTINUOUS VARIABLES INTO CATEGORIES FOR ANALYSIS OF VARIANCE: COUNTERPRODUCTIVITY AND ERRORS

ABSTRACT: The analysis of variance (ANOVA) with categorical factors is nothing more than a special case of ANOVA with continuous factors that is usually called regression analysis. Simulations with Monte Carlo methods (MMC) were performed with the aim of showing that the conversion of continuous variables into categories (common procedure for tests with wood) is almost always counterproductive and often inducing error. In one experiment, where the percentage of particles in a particleboard produced with bamboo leaves and particles was converted into five equidistant categories between 0% and 100%, the parallel tension to each category was obtained. The linear model obtained in ANOVA with continuous factors was simulated in seven different conditions of variability (coefficients of variation between 1% and 20%). The results indicate that the approach of converting continuous variables into categories, although attractive, can lead to misleading conclusions: the same data as in ANOVA with factors produced continuous linear models where the coefficient of determination (R^2) was low and therefore inadequate to describe efficiently the observed variability, produced a high percentage of significant multiple comparisons (Tukey; $\alpha = 0.05$) in ANOVA with categorical factors. For instance, the same data that produced an average coefficient of determination close to 0.50 yield on average 50% of significant multiple comparisons.

Keywords: simulation, Monte Carlo methods (MCM), parallel tension, ANOVA

1. Introdução.

O modelo linear usado para representar os valores observados de um experimento inteiramente casualizado é:

$$y_{ij} = \mu + \alpha_i + e_{ij} \quad (1)$$

onde y_{ij} representa a resposta correspondente a j-ésima no i-ésimo tratamento; μ a média geral, α_i é o efeito devido ao i-ésimo nível do fator qualitativo (categórico) α e e_{ij} é o erro devido à ij-ésima parcela.

Quando o fator α for **quantitativo** (contínuo) o modelo pode ser reescrito na forma:

$$y_{ij} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \dots + \beta_{a-1} X_i^{a-1} + e_{ij} \quad (2)$$

onde X_i representa os níveis do fator α . É possível determinar o componente de regressão linear (β_1), quadrático (β_2), cúbico (β_3), etc., até o (a-1)-ésimo grau, (β_{a-1}). Neste modelo β_0 é o parâmetro linear e corresponde a μ do modelo (1) enquanto que o efeito de tratamentos representado por α_i está agora particionado nos componentes de regressão representados por β_1, β_2 , etc.

Na literatura estatística a análise de variância do modelo (1) é denominada como ANOVA, enquanto que a análise de variância do modelo (2) é denominada como análise de regressão.

“Basicamente, quando um pesquisador categoriza uma variável contínua em nome do delineamento experimental, está trocando uma inferência fraca a respeito de uma categoria forte (há apenas uma categoria) por uma inferência forte a respeito de categorias subjetivas, e portanto fracas” (MAGNUSSON e MOURÃO, 2005). A conversão de variáveis contínuas em categorias, uma abordagem que ocorre com frequência nos trabalhos de ciência e tecnologia da madeira, quase sempre é contraproducente e frequentemente, enganadora.

2. Material e Métodos.

Os estudos de simulação tentam reproduzir num ambiente controlado as situações que ocorrem com um problema real. Uma simulação pode ser entendida como uma particular realização de um modelo (binomial, normal, etc.) e nesse sentido, os valores simulados podem ser considerados como uma amostra. A simulação de variáveis aleatórias deu origem aos métodos Monte Carlo que pressupõem a utilização de geradores de números pseudoaleatórios.

O gerador de números pseudoaleatórios utilizado foi o Mersenne twister, desenvolvido em 1997 por Makoto Matsumoto e Nishimura Takuji, que é baseado em recorrência linear matricial sobre um campo binário (um campo que contém dados na forma de números binários) finito F2. Esse algoritmo foi projetado especificamente para corrigir muitas das falhas encontradas em algoritmos mais antigos e gera rapidamente números pseudoaleatórios de altíssima qualidade. Seu nome tem origem no fato de que a duração do período deve ser um número de Mersenne ($M_p = 2^p - 1$) primo.

Há pelo menos duas variantes do algoritmo comum, que diferem apenas no tamanho dos primos de Mersenne usados: o MT19937 (o mais recente e mais utilizado) com comprimento de palavra de 32 bits e o MT19937-64 com comprimento de palavra de 64 bits. O Mersenne twister gera números com uma distribuição uniforme no intervalo $[0, 2^k - 1]$, onde k é o número de bits da palavra.

O Mersenne twister é amplamente utilizado. É o gerador de números pseudoaleatórios padrão no R (que foi a linguagem de programação e o ambiente para computação estatística utilizado neste trabalho na sua versão **R 2.10.1**), no Maple, no Matlab e nas duas linguagens de script mais populares, Python e Ruby. Ele é particularmente indicado para simulações com