Uma aplicação em delineamentos inteiramente casualizados com a distribuição assimetrica t-Student

Altemir da Silva Braga [1], Raquel Aline Oliveira Eloy [2]

[1] Doutorando em Estatística e Experimentação Agronômica / ESALQ.
[2] Mestranda em Estatística e Experimentação Agronômica / ESALQ.

RESUMO

Apresentar um novo estudo na área de estatística com aplicações a dados reais sempre será um desafio para os pesquisadores, principalmente, da área de estatística experimental. Neste, trabalho utilizou-se a distribuição assimétrica t-Student assimétrica tipo 3 (ST3) com quatro parâmetros para avaliar o efeito do teor de B e a absorção de S na produção de grãos de soja. Essa distribuição é simétrica, assimétrica, platicúrtica, leptocúrtuca, unimodal e bimodal assimétrica para alguns valores paramétricos. Encontra-se bem definida e fundamentada por meio de propriedades matemáticas. As estimativas dos parâmetros foram obtidas utilizando o método da máxima verossimilhança. Foram realizados estudos de simulação para diferentes cenários e, ainda, a análise de resíduos. O novo estudo obteve melhores resultados em relação ao modelo normal, os conforme os critérios de comparação de modelos AIC BIC.

Palavras chave: Delineamentos; Assimetria; Estimação; Verossimilança; Comparação; Resíduos.

1 INTRODUÇÃO

A estatística experimental é a ferramenta mais indicada para trabalhar com dados provenientes de ensaios experimentais. Ela vem contribuindo na pesquisa científica desde o planejamento até a interpretação dos resultados. Este fato que faz com que a análises estatísticas desempenhem um papel fundamental no campo científico, visto que tais técnicas são utilizadas em quase todas as etapas da pesquisa. Neste contexto, muitos estudos são realizados na estatística experimental com a finalidade de melhorar as análises e ajudar na interpretação dos resultados.

Neste contexto, realizou-se este estudo utilizando a distribuição t de Student assimétrica tipo 3 que encontra-se no pacote "GAMLSS"do software R. Este método faz parte de uma ampla classe de modelos estatísticos chamada GAMLSS (Modelos Aditivos Generalizados para Posição, Escala e Forma). Tal distribuição foi proposta e fundamentada matematicamente utilizando o método de (FERNÁNDEZ C.; STEEL, 1995). Esta metodologia destaca-se, pois não requer a função de distribuição acumulada. Além disso, propriedades matemáticas como os momentos são fáceis de serem obtidos. Na literatura encontram-se aplicações em retornos financeiros, em (FERNÁNDEZ C., 1998), entre outras aplicações, ver (AZZALINI; CAPITANIO, 2003).

Dessa forma, conduziu-se este estudo, com o objetivo de ajustar o modelo de regressão na distribuição assimétrica t-Student tipo 3. O ajuste do modelo verificou

o efeito do teor de boro e a absorção de enxofre na produção de grãos de soja e, além disso, foram comparados os modelos normal e ST3 utilizando os critérios AIC e BIC. Dessa forma, após as análises estatísticas, a distribuição assimétrica ST3 ajustou-se melhor ao conjunto de dados em relação à distribuição normal.

2 Materiais e métodos

Os dados utilizados nesta aplicação são referentes a um ensaio experimental conduzido no Bloco 3, do Centro de Pesquisa Geraldo Schultz, localizado no município de Iracemápolis-SP. O clima segundo Köppen é do tipo Cwa (clima tropical de altitude, com chuvas no verão e seca no inverno). O solo foi classificado como Latossolo Vermelho distrófico. O objetivo do experimento foi avaliar o efeito do teor de boro (B) e a absorção de enxofre (S) na produção de grãos de soja. A pesquisa foi realizada pela empresa Produquimica, em Iracemápolis-SP, no ano agrícola 2014/2015.

A semeadura foi realizada no dia 12 de novembro de 2014, com espaçamento entre linhas de 0,5 m e densidade de 16 sementes por metro linear. Cada parcela foi composta de 6 linhas com 7 m de comprimento, com 6 repetições por tratamento. A parcela útil foi constituída por 2 linhas com 5 m de comprimento. Foi utilizado um tratamento controle (Tratamento 1) e, além disso, o experimento foi conduzido por meio de um delineamento inteiramente casualizado com 4 repetições e 7 tratamentos.

3 Estimação por máxima verossimilhança

Seja $y_{11}, \ldots, y_{IJ}$ uma amostra de tamanho n da distribuição OLLSN. Então, considerando o logarítmo da função verossimilhança da distribuição marginal, supondo o modelo normal para o efeito aleatório e o vetor de parâmetros $\boldsymbol{\theta} = (m, \boldsymbol{\tau}^T, \sigma, \lambda, \alpha)^T$, em que $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_I)^T$. Tem-se que o logaritmo da função de verossimilhança da expressão pode ser reescrita da eguinte forma:

$$Y_{ij} = m + \boldsymbol{\tau}_i + \sigma\epsilon_{ij}, \tag{1}$$

em que Y_{ij} representa o valor observado do tratamento i, m é o efeito da média geral, τ_i é o efeito do tratamento i e $\epsilon_{ij} \sim \text{ST3}(0, \sigma^2, \nu, \alpha)$ representam os efeitos do fator não controlado do ensaio experimental, com $i = 1, ..., I$ e $j = 1, ..., J$, sendo que I denota o número de tratamentos e J o número de repetições.

Seja $y_{11}, ..., y_{IJ}$ uma amostra de tamanho n da distribuição $ST3$. Então, o logaritmo da função verossimilhança para o vetor de parâmetros $\boldsymbol{\theta} = (m, \boldsymbol{\tau}^T, \sigma, \tau, \nu)^T$, em que $\boldsymbol{\tau} = (\tau_1, ..., \tau_I)^T$ é dada por:

$$l(\boldsymbol{\theta}) \;=\; \sum_{i=1}^{I}\sum_{j=1}^{J} log\left\{k\sigma^{-1} \times \left[\left(1 + \frac{z^2}{\nu}\right)\frac{1}{\tau^2}\right]^{-\frac{(\nu+1)}{2}}\right\} \tag{2}$$

em que, $z_{ij} = (y_{ij} - m - \tau_i)/\sigma$. As estimativas de máxima verossimilhança $\widehat{\boldsymbol{\theta}}$ do vetor de parâmetros podem ser obtidas maximizando a log-verossimilhança (2). Nesta etapa, foram utilizados os métodos Nelder-Mead e o "L-BFGS-B" que são fornecidos

no pacote "Optim" do *software* R, em que os os valores iniciais podem ser obtidos da função "summary.lm". Além disso, o "Optim" fornece a opção "gr" em que o usuário pode fornecer o vetor escore que torna o algoritmo mais eficiente. Neste sentido, para obter as componentes do vetor escore $U(\boldsymbol{\theta})$ do modelo ST3 derivou-se a função log-verossimilhança (2) em relação a ν, m, μ, $\boldsymbol{\tau}$ e σ. Dessa forma, as componentes do vetor escore $U(\boldsymbol{\theta})$ são dadas por:

$$\begin{aligned}
U_{\nu}(\boldsymbol{\theta}) &= \sum_{i=1}^{I}\sum_{j=1}^{J}\left\{\frac{\psi\left(\frac{\nu+1}{2}\right)\sqrt{\pi\nu}\boldsymbol{\tau}^2+\Gamma\left(\frac{\nu}{2}\right)\left[\psi\left(\frac{\nu+1}{2}\right)\pi\nu-\psi\left(\frac{\nu}{2}\right)\pi\nu-\pi\right]}{2\sqrt{\pi\nu}}\right\} \\
&\quad +\sum_{i=1}^{I}\sum_{j=1}^{J}\left\{-\frac{1}{2}ln\left[\frac{\nu\sigma^2+(y_{ij}-\mu)^2}{\nu\sigma^2\boldsymbol{\tau}^2}\right]+\frac{(\nu+1)+(y_{ij}-\mu)^2}{2\nu\left[\nu\sigma^2+(y_{ij}-\mu)^2\right]}\right\}, \\
U_{\mu}(\boldsymbol{\theta}) &= \sum_{i=1}^{I}\sum_{j=1}^{J}\left\{\frac{\left(\frac{\nu+1}{2}\right)(-2y_{ij}+2\mu)}{\nu\sigma^2+(y_{ij}-\mu)^2}\right\}, \\
U_{\boldsymbol{\tau}}(\boldsymbol{\theta}) &= \sum_{j=1}^{J}\left\{\frac{-\boldsymbol{\tau}^2+\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}-(\nu+1)\left[\boldsymbol{\tau}^2+\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}\right]}{\boldsymbol{\tau}\left(\boldsymbol{\tau}^2+\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}\right)}\right\}, \\
U_{\sigma}(\boldsymbol{\theta}) &= \sum_{i=1}^{I}\sum_{j=1}^{J}\left\{\frac{1}{\sigma}-\frac{(\nu+1)(y_{ij}-\mu)^2}{\sigma[\nu\sigma^2+(y_{ij}-\mu)^2]}\right\}.
\end{aligned}$$

Igualando essas equações a zero e resolvendo-as, simultaneamente, obtêm-se as estimativas de máxima verossimilhança dos parâmetros utilizando métodos numéricos.

Sob certas condições de regularidade o vetor de parâmetros $\boldsymbol{\theta}$, em seu espaço paramétrico, tem distribuição assintótica $\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ normal multivariada $N_{I+4}(0, K(\boldsymbol{\theta})^{-1})$, em que $K(\boldsymbol{\theta})$ é a matrix de informação esperada. A distribuição normal assintótica $N_{I+4}(0, -\ddot{\mathbf{L}}(\boldsymbol{\theta})^{-1})$ pode ser utilizada para construir regiões aproximadas de confiança para o vetor de parâmetros $\boldsymbol{\theta}$. Neste caso, os intervalos de confiança assintóticos $100(1-\gamma)\%$ para cada componente do vetor de parâmetros θ_r é dado por:

$$IC_r = \left(\hat{\theta}_r - z_{\gamma/2}\sqrt{-\hat{\ddot{L}}^{r,r}}, \hat{\theta}_r + z_{\gamma/2}\sqrt{-\hat{\ddot{L}}^{r,r}}\right),$$

em que $-\hat{\ddot{L}}^{r,r}$ denota o r-ésimo elemento da diagonal da inversa da matriz de informação observada $-\ddot{\mathbf{L}}(\widehat{\boldsymbol{\theta}})^{-1}$ e $z_{\gamma/2}$ é o quantil $1-\gamma/2$ da distribuição normal padrão.

4 Resultados

Na Tabela 1 estão as estimativas dos parâmetros para os modelos normal e ST3. Dentre tais estimativas de máxima verossimilhança (EMVs) destacam-se os erros-padrão e os valores do Critério de Informação de Akaike (AIC) e Critério de Informação Baysiano (BIC). De acordo com os critérios, quanto menor forem estas estimativas, melhor o ajuste. Enquanto que na Tabela 2 encontram-se os valores do teste de razão de verossimilhança e do seu nível descritivo. Conforme os resultados,

existem evidências estatísticas que o modelo ST3 pode ser utilizado para análise dos dados de densidades.

Tabela 1: Estimativas de máxima verossimilhança e os critérios de informação

	Normal				ST3		
$\boldsymbol{\theta}$	MLE	E.P.	p-valor	$\boldsymbol{\theta}$	MLE	E.P.	valor-p
m	23.400	0.986	<0.001	m	22.617	0.186	<0.001
τ_2	0.625	1.395	0.659	τ_2	-1.204	0.302	<0.001
τ_3	-1.250	1.395	0.380	τ_3	-0.796	0.246	<0.001
τ_4	4.125	1.395	0.007	τ_4	3.586	0.497	<0.001
τ_5	-1.625	1.395	0.257	τ_5	-1.008	0.350	<0.001
τ_6	-0.150	1.395	0.915	τ_6	0.001	0.254	0.994
τ_7	0.000	1.395	1.000	τ_7	0.199	0.271	0.472
σ	0.679	0.133	<0.001	σ	0.204	0.266	<0.001
ν				ν	2.372	0.187	<0.001
α				α	0.881	0.324	0.701
	DG	AIC	SBC		DG	AIC	SBC
	117.512	133.512	144.169		79.765	99.765	113.087

Tabela 2: Teste de razão de verossimilhança para verificar se existe diferença estatística entre os modelos ST3 e o normal.

Modelos	Hipóteses	Estatística w	valor-p
Normal vs ST3	$H_0 : \lambda = 1$ vs $H_1 : H_0$ é falsa	37,747	$< 0{,}001$

5 Considerações finais

A nova distribuição de probabilidade com quatro parâmetros chamada assimétrica t-Student assimétrica tipo 3 (ST3) foi proposta com sucesso nos estudos com ensaios de delineamento inteiramente casualizado. O novo modelo ajustou-se melhor do que a distribuição normal.

Referências

[1] FERNÁNDEZ C., O. J.; STEEL, M. F. J. Modeling and inference with v- spherical distribution. Journal of the American Statistical Association, n. 90, 1995.

[2] FERNÁNDEZ C., S. M. F. J. On bayesian modeling of fat tails and skewness. Journal of the American Statistical Association, v. 93, n. 441, 1998.

[3] AZZALINI, A.; CAPITANIO, A. Distribuitions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. v. 65, 2003.