模型选择信息量准则 *AIC* 及其在方差分析中的应用

宋喜芳,李建平,胡希远
(西北农林科技大学 农学院,陕西 杨凌 712100)

[摘 要] 【目的】探讨模型选择信息量准则 *AIC* 在方差分析模型选择中的必要性和意义。【方法】简要介绍了模型选择信息量准则 *AIC* 的概念,推导了 *AIC* 在方差分析模型选择的公式,并运用 *AIC* 对水稻品种比较试验数据进行了最佳 ANOVA 模型选择分析。【结果】对 5 个水稻品种比较试验进行方差分析模型的选择是必要的;*AIC* 准则在方差分析模型选择中是一种简单有效的方法。【结论】*AIC* 可用于方差分析模型的选择,以改进方差分析结论的可靠性。

[关键词] 方差分析;*AIC*;品种比较试验;模型选择
[中图分类号] O212.1 [文献标识码] A [文章编号] 1671-9387(2009)02-0088-05

Model selection criterion *AIC* and its application in ANOVA

SONG Xi-fang, LI Jian-ping, HU Xi-yuan
(*College of Agriculture, Northwest A&F University, Yangling, Shaanxi* 712100, *China*)

Abstract: 【Objective】 The aim of this study was to conduct the necessity and significance of the model selection criterion *AIC* in ANOVA selection. 【Method】 This study introduced the conception of model selection criterion *AIC* briefly, deduced *AIC* formula in ANOVA model selection and applied *AIC* to select the best ANOVA model in rice variety comparison trail data. 【Result】 ANOVA model selection was necessary in the 5 rice varieties comparison trial data. *AIC* was a simple and effective method in ANOVA model selection. 【Conclusion】 *AIC* could be used to select the best model of ANOVA and improve the credibility of the ANOVA conclusion.

Key words: ANOVA; *AIC*; genotype comparison trial; model selection

模型是人们对客观现象或过程认识的反映和描述,建立模型则是分析研究有关问题的方法。模型选择在数据分析中具有重要意义[1]。最佳模型选择已成为许多科学研究的中心问题和各种统计分析的重要基础[2]。要获得可靠的研究结论,除具备合理的调查或试验设计、严格的试验操作及可靠的数据等条件外,科学合理地分析模型或方法也起着非常重要的作用,这在通常需要对模型变量和结构形式进行筛选的各种回归分析中得到充分体现,并被人们普遍认识、重视和应用。方差分析(ANOVA)是工农业生产和科学研究中,对试验数据进行分析的一种重要的数理统计方法,其应用极其广泛。但是,受习惯、可供应用统计方法和分析软件等因素的限制,传统上通常是依据一定试验设计中试验因子、环境因子及其之间互作等可能效应的多少来确定 ANOVA 模型,极少考虑针对特定试验数据的最佳 ANOVA 模型选择。即,不管一些因子效应在具体试验中是否真实存在,只要试验设计给定,则其

* [收稿日期] 2008-03-31
[基金项目] 国家自然科学基金项目(30571072);教育部留学回归人员基金项目(2005-2007)
[作者简介] 宋喜芳(1982-),女(壮族),广西南宁人,在读硕士,主要从事生物统计研究。
[通信作者] 胡希远(1963-),男,陕西蓝田人,副教授,博士,主要从事生物统计应用研究。E-mail:xiyuanhu@yahoo.com.cn

ANOVA 模型便确定。这样的 ANOVA 模型未必就一定能最佳体现所有试验数据的信息;即使偶尔有考虑 ANOVA 模型的选择问题,也多是采用传统的具有一定局限性的 F 检验结果进行因子选择[3]。因此,传统方差分析法分析结果的可靠性均受到不同程度的限制。为此,本文介绍了一种在统计分析,特别是在统计模型选择中广泛应用的赤池信息量准则(Akaike's Information Criteria,常简称 *AIC*),推导其在 ANOVA 模型选择中的具体公式,并用具体试验分析实例证明 *AIC* 遴选品种试验最佳 ANOVA 模型的效果,以期为 *AIC* 在方差分析中的应用,及提高试验数据分析的准确性和研究结论的可靠性提供依据。

1　*AIC* 准则的定义

为了理解 *AIC* 的原理和定义,需要了解 Kullback-Leibler 信息。假设存在一个真模型,统计分析和选择的目标就是要从一系列侯选模型中选出和真模型最接近的模型。在模型选择中,人们希望选到最佳的模型。依据信息理论,如果真模型和一个待选模型的概率密度函数分别为 $f(x,\theta^*)$ 和 $g(x,\theta)$,则两模型间的距离可用 Kullback-Leibler 信息来表示:

$$I=(f(x,\theta^*),g(x,\theta))=\int f(x,\theta^*)\ln\frac{f(x,\theta^*)}{g(x,\theta)}\mathrm{d}x=$$

$$\int f(x,\theta^*)\ln f(x,\theta^*)\mathrm{d}x-\int f(x,\theta^*)\ln g(x,\theta)\mathrm{d}x。$$

该式也被称为 Kullback-Leibler 距离,或简称 K-L 距离。其中 x 是 n 个相互独立的样本观测数据。可以证明,$I=(f(x,\theta^*),g(x,\theta))\geqslant 0$ 和当且仅当 $f(x,\theta^*)=g(x,\theta)$ 时,I 等于 0。$I=(f(x,\theta^*),g(x,\theta))$ 值越小,$g(x,\theta)$ 越接近 $f(x,\theta^*)$。在 K-L 距离中,第 1 项虽是不可估的,但对所有待选模型是相同的,因此可不考虑;第 2 项则可以从给定的样本数据估计。这一策略被称作熵最大原理(EMP)[4]。在统计推论中,采用不同的方法估计第 2 项可得到不同的信息准则。这些信息准则中较重要的就是日本学者赤池弘次(Akaike)在研究信息论,特别是在解决时间序列定阶问题时提出来的 *AIC* 准则。他将 K-L 距离和极大似然方法相结合,并利用似然估计性质,推导出了最佳模型选择的 *AIC* 准则,其定义为[5]

$$AIC=-2\ln(\text{极大似然函数})+2k。\qquad(1)$$

式中:第 1 项是极大似然函数自然对数值乘以 −2,第 2 项是模型中独立参数维数 k 的 2 倍。赤池弘次建议,当从 1 组可供选择的模型中选出 1 个最佳模型时,选择 *AIC* 为最小的模型是可取的。当在 2 个模型中存在着相当大的差异时,这个差异出现于式(1)右边第 1 项,而当第 1 项不出现显著性差异时,第 2 项就起作用,从而参数个数少的模型是好的模型。所以,第 2 项可以解释为对增加模型参数的一种惩罚,是衡量模型拟合优度的一个量。因此在理论结构上看,*AIC* 采用了最小限度的定义,具体化地采用了"吝啬原理"。由此看来,数据的拟合既好而又尽可能节省参数数目的模型才是最佳的。

2　*AIC* 在 ANOVA 模型的推导

考虑到普通农业科学试验数据分析者更习惯于应用 ANOVA 平方和分解的方法估计参数[6],而极大似然估计是一种相对新的参数估计方法,本文将推导 *AIC* 在 ANOVA 中的公式,以促进 *AIC* 在农业试验数据分析中的推广应用。

ANOVA 模型是一类带有线性约束的统计模型,一般数学表达式为

$$y_i=\beta_0+\sum_{j=1}^{p}x_{ij}\beta_j+\varepsilon_i,i=1,2,\cdots,n;j=1,2,\cdots,p。\qquad(2)$$

式(2)的未知参数是因子效应。试验是控制在条件基本相同的情况下进行,不论是单因素方差分析还是多因素方差分析,尽管各处理均值 μ_i 可能会有较大差异,但认为误差 ε_i 相互独立、方差同质服从正态分布[7],即 $\varepsilon_i\sim N(0,\sigma^2)$,假设 $y_i\sim N(\mu_i,\sigma^2)$。这时有

$$\mu_i=E(y_i)=E(\beta_0+\sum_{j=1}^{p}x_{ij}\beta_j+\varepsilon_i)=\beta_0+\sum_{j=1}^{p}x_{ij}\beta_j。$$

从而有 $y_i\sim N(\beta_0+\sum_{j=1}^{p}x_{ij}\beta_j,\sigma^2)$,可得到 y_i 的分布密度函数

$$f(y_i)=\frac{1}{\sqrt{2\pi}\sigma}\exp(\frac{-1}{2\sigma^2}(y_j-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j)^2)。$$

由似然函数的定义得其似然函数为

$$L=L(y_i;\theta)=L(y_i;\beta_0,\beta_j,\sigma^2)=$$

$$\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp(\frac{-1}{2\sigma^2}(y_j-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j)^2)=$$

$$(2\pi\sigma^2)^{\frac{-n}{2}}\exp(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_j-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j)^2)。\qquad(3)$$

对式(3)两边取对数,得

$$\ln L=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-$$

$$\sum_{j=1}^{p}x_{ij}\beta_j)^2。\qquad(4)$$

对式(4)中的参数进行极大似然估计,也就是对参数分别进行求偏导数,并令其为零,得:

$$\begin{cases}\dfrac{\partial \ln L}{\partial \beta_0}=0,\\ \dfrac{\partial \ln L}{\partial \beta_{j_0}}=0,\\ \dfrac{\partial \ln L}{\partial \sigma^2}=0。\end{cases}$$

其中:j_0 为满足 $1\leqslant j_0\leqslant p$ 的任意整数。

解该方程组,可得 β_0,β_{j_0} 和 σ^2 的极大似然估计量 $\hat{\beta}_0$,$\hat{\beta}_{j_0}$,$\hat{\sigma}^2$ 如下:

$$\hat{\beta}_0=\frac{1}{n}\sum_{i=1}^{n}(y_i-\sum_{j=1}^{p}x_{ij}\hat{\beta}_j),\tag{5}$$

$$\hat{\beta}_{j_0}=\frac{1}{\sum_{i=1}^{n}x_{ij_0}^2}\sum_{i=1}^{n}x_{ij_0}[(y_i-\hat{\beta}_0-(\sum_{j=1,j\neq j_0}^{p}x_{ij}\hat{\beta}_j))],\tag{6}$$

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\sum_{j=1}^{p}x_{ij}\hat{\beta}_{j_0})^2。\tag{7}$$

由式(7)有:$\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\sum_{j=1}^{p}x_{ij}\hat{\beta}_{j_0})^2=n\hat{\sigma}^2$,代入式(4),于是模型极大似然函数自然对数为:

$$\ln L=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln\hat{\sigma}^2-\frac{n}{2}。\tag{8}$$

把式(8)代入式(1)得到

$$AIC=n\ln(2\pi)+n\ln\hat{\sigma}^2+n+2p,$$

$$AIC=n\ln\hat{\sigma}^2+2p+c。$$

此式即为 ANOVA 模型中的 AIC。其中,$c=n\ln(2\pi)+n$ 与模型中独立参数维数 p 无关,是对所有候选模型相同的常数,比较时可忽略,即方差分析模型选择实际上可用:

$$AIC=n\ln\hat{\sigma}^2+2p。\tag{9}$$

3 AIC 在 ANOVA 模型的应用实例

本文以文献[8]中的试验数据为例。该试验为 A、B、C、D、E 5 个水稻品种的比较试验,E 是标准品种,采用 拉丁方设计,其田间排列和产量观测结果见表 1。

表 1 水稻品种试验拉丁方设计的田间排列和水稻产量

Table 1 Field arrangement and Paddy yield of Latin design

横行区组 Row block	纵行区组 Column block				
	1	2	3	4	5
1	D(37)	A(38)	C(38)	B(44)	E(38)
2	B(48)	E(40)	D(36)	C(32)	A(35)
3	C(27)	B(32)	A(32)	E(30)	D(26)
4	E(28)	D(37)	B(43)	A(38)	C(41)
5	A(34)	C(30)	E(27)	D(30)	B(41)

注:括弧中数字为各品种在对应行列的产量,单位为 kg。

Note: The umbers in bracket is each variety's yield according to its row and column, with kg as the unit.

对应该试验设计通常采用的方差分析模型为

$$y_{ijk}=\mu+V_i+R_j+C_k+e_{ijk}。\tag{10}$$

式中:μ 为总体平均值,V_i 为第 i($i=1,2,3,4,5$)水稻品种的效应,R_j 为第 j($j=1,2,3,4,5$)行区组的效应,C_k 为第 k($k=1,2,3,4,5$)列区组的效应,y_{ijk} 和 e_{ijk} 分别为 i 品种在 j 行 k 列的产量观测值和对应的误差。依据该方差分析模型,得到的方差分析结果如表 2 所示。从表 2 中 F 检验结果可知,品种效应达显著水平,行区组效应达极显著水平,而列区组效应不显著。通常情况下方差分析到此便做出统计结论。然而,列区组效应统计检验不显著,则意味着列区组效应不存在,其离均差平方和 $SS=6.64$ 仅是由误差引起,该离均差平方和应合并到误差项中。把列区组的离均差平方和合并到误差项后,得到的方差分析结果如表 3 所示。

表 2 水稻品种试验传统方差分析方法的结果

Table 2 Results of the traditional ANOVA

变异来源 Variation resource	自由度 df	离均平方和 SS	均方差 MS	F 检验 F-test	概率值 Pr
品种 Variety(V)	4	271.44	67.86	4.32	0.021 5*
行 Row(R)	4	348.64	87.16	5.55	0.009 1**
列 Rank (C)	4	6.64	1.66	0.11	0.978 3
误差 Error(E)	12	188.32	15.693		

注:* 表示在 $\alpha=0.05$ 水平显著,** 表示在 $\alpha=0.01$ 水平极显著。下表同。

Note: * denotes the significant level of $\alpha=0.05$, ** denotes the most significant level of $\alpha=0.01$.

表 3　水稻品种试验列区组合并到误差项后的方差分析结果

Table 3　Results of ANOVA after combination column block into the error

变异来源 Variation resource	自由度 df	离均平方和 SS	均方差 MS	F 检验 F-test	概率值 Pr
品种 Variety(V)	4	271.44	67.86	5.57	0.005 3**
行 Row(R)	4	348.64	87.16	7.15	0.001 7**
误差 Error($E'=E+C$)	16	194.96	12.185		

由表 3 可以看出,此时品种效应和行区组效应 F 检验的概率较前都发生了变化,均达到了极显著水平。从模型选择的观点来看,表明原给定的方差分析模型(10)不是该试验数据的最佳模型,而从中剔除列区组 C_k 后的模型才是最佳的。由此可见,方差分析也存在最佳模型选择问题[9],模型的选择关系着试验研究的结论。

用式(9)的 AIC 准则来考虑模型选择的问题。该试验观测值数为 $n=5\times5=25$,除剩余误差外,有 3 个构成模型的效应:一个是品种效应,另两个为行区组和列区组效应,则其所有可能的候选模型数为 $C_3^1+C_3^2+C_3^3=7$。但由于品种是试验特别要检验的因素,模型中无品种效应无意义,这样实际上有意义的待选模型有 4 个,一个是模型(10),为方便称之为 M_{VRC},其余为 M_{VRC} 去掉行、列区组效应后形成的 3 个模型,设这些模型的名称分别为 M_{VR}、M_{VC} 和 M_V,该模型构成如下

$$M_{VR}: y_{ijk}=\mu+V_i+R_j+e_{ijk},$$
$$M_{VC}: y_{ijk}=\mu+V_i+C_k+e_{ijk},$$
$$M_V: y_{ijk}=\mu+V_i+e_{ijk}。$$

依据这些模型分别进行方差分析得到有关统计量,如表 4 所示。由表 4 可知,模型 M_{VRC} 的 AIC 不是最小,不是最佳模型。如果按传统方差分析不针对具体的试验数据进行模型选择,其分析结果的准确性必然受到影响。模型 M_{VR} 的 AIC 值最小,即该模型能最佳地反映所分析试验数据的信息,应利用该模型对其试验数据进行分析。这和前述 F 检验选择模型的结果是相同的,表明 AIC 在模型选择上是有效的。

表 4　各种待选模型的统计量

Table 4　Several statistical variables of the candidate model

模型 Model	n	$\hat{\sigma}^2$	$\ln\hat{\sigma}^2$	p	AIC
M_{VRC}	25	188.32	5.238 1	12	154.95
M_{VR}	25	194.96	5.272 8	8	<u>147.82</u>
M_{VC}	25	536.96	6.285 9	8	173.15
M_V	25	543.60	6.298 2	4	165.46

注:标有下划线的数据表示最佳模型。

Note: The best model is underlined.

各种待选模型的 AIC 计算式为:

$$AIC(M_{VRC})=25\ln(188.32)+2\times(4+4+4)=154.95;$$
$$AIC(M_{VR})=25\ln(6.64+188.32)+2\times(4+4)=147.82;$$
$$AIC(M_{VC})=25\ln(348.64+188.32)+2\times(4+4)=173.15;$$
$$AIC(M_V)=25\ln(348.64+6.64+188.32)+2\times4=165.46。$$

4　结论与讨论

不同模型对试验数据信息反映的准确性不同。只有对数据拟合度好而又简单(包括模型参数数目与形式)的模型才是最佳模型[10]。AIC 准则不仅具有严格的理论基础,而且也符合模型选择的"吝啬原理",在数学计算上简单,因此被广泛应用于经济学、心理学、自动控制和生物信息学中的模型研究[11]。本文应用实例表明,方差分析中模型的选择是必要的,AIC 的应用是有效的,此方面在其他文献亦有类似报道[12]。

AIC 不需任何统计表和主观上的议论,例如显著性水平 α 取多大,以及当多个因子检验不显著时如何确定最佳模型。AIC 的利用蕴涵着大量数据统计分析自动化的可能性[13]。本文为方便对 AIC 的理解,逐步对 AIC 进行了计算。实际上许多统计软件,如 SAS 和 SPSS 等中已提供了 AIC 等信息量指标,在数据分析时自动给出各模型的 AIC 值,可直接用于模型选择,应用方便。AIC 不仅可用于方差分析模型的选择,而且可用于传统方差分析不能有效处理的相关数据,如空间相关数据协方差结构

模型的选择[14],在试验数据分析方面具有广泛的应用前景。建议有关试验分析者将 *AIC* 应用于方差分析和其他更复杂试验数据分析模型的选择,以改善数据分析的效果。

[参考文献]

[1] Little R C, Milliken G A, Stroup W W, et al. SAS system for mixed models [M]. Cary, NC: SAS Institute, 1996: 303-326.

[2] Burnham K P, Anderson D R. Model selection and inference: A practical information-theoretical approach [M]. New York: Spring-verlag, 2002: 96-97.

[3] 张 翔,陈建能. 关于回归方程自变量的选择 [J]. 福建农业大学学报,1999,28(3):357-360.
Zhang X, Chen J N. Selection of variables in regression equations [J]. Journal of Fujian Agricultural University, 1999, 28(3): 357-360. (in Chinese)

[4] Akaike H. Stochastic theory of minimal realization [J]. IEEE Trains on Automatic Control, 1974, 19(6): 667-674.

[5] Bozdogan H. Model selection and Akaike's information criterion(AIC): the general theory and its analytical extension [J]. Psychometrika, 1987, 52: 345-570.

[6] 袁志发,周静芋. 多元统计分析[M]. 北京:科学出版社,2002:54-58.
Yuan Z F, Zhou J Y. The analysis of multiple statistic [M]. Beijing: Science Republishing Company, 2002: 54-58. (in Chinese)

[7] 刘光祖. 概率论与数理统计 [M]. 北京:高等教育出版社,2000:291-292.
Liu G Z. Probability theory and symbolic statistic [M]. Beijing: Higher Education Publishing Company, 2000: 291-292. (in Chinese)

[8] 胡希远. SAS与统计分析 [M]. 陕西杨凌:西北农林科技大学出版社,2007:107-108.
Hu X Y. SAS and statistical analysis [M]. Yangling, Shaanxi: Journal of Northwest A&F University Publishing Company, 2007: 107-108. (in Chinese)

[9] 刘璋温. 赤池信息量准则 AIC 及其意义[J]. 数学的实践和认识,1980,3(1):64-72.
Liu Z W. Akaike information criterion and its meaning [J]. Mathematics In Practice and Theory, 1980, 3(1): 64-72. (in Chinese)

[10] 段晓君,王正明. 基于选择准则的参数模型评价方法 [J]. 国防科技大学学报,2003,25(3):62-65.
Duan X J, Wang Z M. Parametric model evaluation based on the selection criterion [J]. Journal of National University of Defense Technology, 2003, 25(3): 62-65. (in Chinese)

[11] 刘璋温,吴国富. 选择回归模型的几个准则 [J]. 数学的实践与认识,1983,1(3):63-71.
Liu Z W, Wu G F. Several criterions of selecting regression model [J]. Mathematics in Practice and Theory, 1983, 1(3): 63-71. (in Chinese)

[12] 王艳君,刘 群,任一屏. AIC 与 BIC 在亲体-补充量模型选择中的应用及比较[J]. 中国海洋大学学报,2005,35(3):397-403.
Wang Y J, Liu Q, Ren Y P. Comparision of AIC and BIC in the Selection of Stock-Recruitment Relationships [J]. Journal of Ocean University of Qingdao, 2005, 35(3): 397-403. (in Chinese)

[13] Thomas M L, Smart L B, Lewis B S. Comparison of the Akaike information criterion, the Schwarz criterion and *F* test as guides to model selection [J]. Journal of Pharmacokine and Biopharmaceutics, 1994, 22(5): 431-445.

[14] 胡希远, Joachim Spilke. 田间试验的空间变异性及其统计控制 [J]. 作物学报,2007,33(4):620-624.
Hu X Y, Joachim Spilke. Spatial variability and its statistical control in field experiment [J]. Acta Agronomical Sinica, 2007, 33(4): 620-624. (in Chinese)

(上接第87页)

[8] 黄从德,胡庭兴,赖家明. 四川巨桉短周期工业原料林二元材积表的编制 [J]. 四川农业大学学报,2003,21(2):106-108.
Huang C D, Hu T X, Lai J M. The establishment of volume table with two factors for pulpwood plantation of *Eucalyptus grandis* in Sichuan [J]. Journal of Sichuan Agriculture University, 2003, 21(2): 106-108. (in Chinese)

[9] 黄从德,胡庭兴,赖家明. 四川巨桉短周期工业原料人工林直径分布规律及其收获模型的研究[J]. 四川林业科技,2003,24(3):41-45.
Huang C D, Hu T X, Lai J M. Research on the DBH distribution and yield model of pulpwood plantation of *Eucalyptus grandis* in Sichuan Province[J]. Journal of Sichuan Forestry Science and Technology, 2003, 24(3): 41-45. (in Chinese)

[10] Riehards B N, Beige D I. Principles and practices of foliar analysis as a basis for croplogging in pine plantations, Basic Considerations [J]. Plant Soil, 1972, 36: 109-119.

[11] 陈竹君. 一种林木营养诊断法——DRIS 法[J]. 陕西林业科技,1993(2):38-40.
Chen Z J. A diagnosis method of forest nutrient——DRIS[J]. Shaanxi Forest Science and Technology, 1993(2): 38-40. (in Chinese)

[12] 黄从德,胡庭兴,赖家明. 四川巨桉短周期工业原料人工林生长规律研究 [J]. 四川林业科技,2003,24(1):70-74.
Huang C D, Hu T X, Lai J M. An research on the growth laws of *Eucalyptus grandis* in Sichuan province [J]. Journal of Sichuan Forestry Science and Technology, 2003, 24(1): 70-74. (in Chinese)