

PROC MIXED 在多误差试验分析中的应用*

胡希远

(西北农林科技大学 农学院, 陕西 杨凌 712100)

[摘要] 以条区试验设计为例, 讨论了多误差试验数据分析的特点, 以及利用统计软件 SAS 中传统程序 PROC ANOVA 和 PROC GLM 对其进行分析的局限性, 介绍和讨论了 PROC MIXED 对试验数据进行统计分析的原理、特点和应用效果, 在此基础上提出了利用 PROC MIXED 对多误差试验数据分析的建议。

[关键词] 试验分析; PROC GLM; PROC MIXED

[中图分类号] S11⁺ 4

[文献标识码] A

[文章编号] 1671-9387(2005)01-0115-05

农业试验数据的分析以统计模型为基础, 其中许多试验的模型为涉及多个误差效应的混合线性模型^[1]。由于这类试验模型中存在多个误差效应, 其试验因素效应统计测验对应的误差项往往不象单误差试验模型那样一定是剩余均方, 而可能是其他均方或者几个均方的线性组合, 因而这类试验的分析相对较复杂, 在利用统计软件进行分析时尤应注意, 不仅要熟悉其方差分析的过程, 而且还要对所用软件的分析机理有所了解, 这样才能得出正确的研究结论。近几年, 在试验数据分析中应用最为广泛的程序是统计软件 SAS 中的 PROC ANOVA 和 PROC GLM。这 2 个程序进行分析的原理基本相同, 前者只对单误差平衡数据, 而后者对单误差平衡数据及非平衡数据的分析均可自动地给出正确的测验结果, 但它们对多误差试验数据的分析未必一定能给出完全正确的测验结果。SAS 中最新开发的 PROC MIXED 程序具有广泛的适用性, 但关于该程序的介绍和应用国内目前尚未见报道。作者拟对多误差试验分析的特点以及利用 PROC GLM 进行分析的局限性予以探讨, 阐述分析利用 PROC MIXED 分析试验数据的原理方法和优点, 以期对科学合理地利利用统计软件分析试验数据有所帮助。

1 多误差试验的统计特点与利用 PROC GLM 进行分析的局限性

农业试验中常用的裂区设计、再裂区设计和条区设计等, 其统计模型含有 2 个以上的误差, 即所谓

的多误差试验设计。下面以条区设计为例, 说明多误差试验分析的特点以及利用 PROC GLM 对其进行分析的局限性。条区设计的统计模型为^[2]:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + bl_k + e_{ik} + e_{jk} + e_{ijk}$$

$$(i = 1, \dots, a; j = 1, \dots, b; k = 1, \dots, r)$$

式中, μ 为总体平均值; α_i 为因素 A 第 i 水平的效应(固定); β_j 为因素 B 第 j 水平的效应(固定); $(\alpha\beta)_{ij}$ 为因素 A 第 i 水平与因素 B 第 j 水平交互效应(固定); bl_k 为第 k 个区组的效应(随机); e_{ik} 和 e_{jk} 为随机误差效应; e_{ijk} 为剩余误差效应, $e_{ik} \sim N(0, \sigma_{ik}^2)$; $e_{jk} \sim N(0, \sigma_{jk}^2)$; $e_{ijk} \sim N(0, \sigma_{ijk}^2)$

对应于上述模型条区设计的方差分析表和效应多重比较的方差及其自由度如表 1 和表 2 所示。

由表 1 和表 2 可知, 无论是 F 测验还是多重比较 (t 测验), 各效应测验误差的组成不同, 只有选用相应的误差项才能得到正确的测验结果。较为复杂的是, 区组效应 F 测验和 $A \times B$ 交互效应多重比较的误差项是方差分析表 1 中多个均方 MS (E_C , E_A 和 E_B) 的线性组合, 其误差的自由度不能由表 1 中 MS 的自由度来准确确定, 而要通过近似计算才能得到^[3,4]。Milliken 等^[4]建议采用 Satterthwaite 法^[5,6]进行自由度近似计算。根据 Satterthwaite 法的原理, 如果一方差 V 为 n 个均方 MS_i 的线性组合, 即

$$V = \sum_{i=1}^n k_i MS_i$$

则该方差 V 的自由度 df_v 为:

* [收稿日期] 2004-02-23

[基金项目] 西北农林科技大学校长基金资助项目

[作者简介] 胡希远(1963-), 男, 陕西蓝田人, 副教授, 博士, 主要从事农业试验模型与模拟分析研究。

$$df_v = \frac{(\sum_{i=1}^n k M S_i)^2}{\sum_{i=1}^n [(k M S_i)^2 / df_i]} \quad (2)$$

式中, k_i 为常数, df_i 为对应于 $M S_i$ 的自由度。以表2中 $A \times B$ 交互效应同 B 异 A 的比较(以 $A \times B/B$ 表示)为例, 其误差均方 $V_{(A \times B/B)}$ 为均方 E_A 和 E_C 的线性

组合: $V_{(A \times B/B)} = h_A E_A + h_C E_C$, 其中, $h_A = 2(b-1)/(br)$, $h_C = 2/(br)$, 则根据式(2)得到 $V_{(A \times B/B)}$ 的自由度 df 为:

$$df = \frac{(h_A E_A + h_C E_C)^2}{(h_A E_A)^2 / df_A + (h_C E_C)^2 / df_C}$$

其中 $df_A = (a-1)(r-1)$, $df_C = (a-1)(b-1)(r-1)$, 它们分别为 E_A 和 E_C 的自由度(表1)。

表1 条区设计方差分析

Table 1 ANOVA table for strip plot design

变异来源 Sources of variation	自由度 df	均方 MS	期望均方 E(MS)	F
区组 Block	$r-1$	$M S_{B1}$	$\sigma_{rest}^2 + b\sigma_A^2 \times B1 + a\sigma_B^2 \times B1 + Q_{B1}$	$M S_{B1} / (Ea + Eb - Ec)$
A	$a-1$	$M S_A$	$\sigma_{rest}^2 + b\sigma_A^2 \times B1 + r b Q_A$	$M S_A / E_a$
$E_a(A \times b1)$	$(a-1)(r-1)$	E_A	$\sigma_{rest}^2 + b\sigma_A^2 \times B1$	
B	$b-1$	$M S_B$	$\sigma_{rest}^2 + a\sigma_B^2 \times B1 + r a Q_B$	$M S_B / E_b$
$E_b(B \times b1)$	$(b-1)(r-1)$	E_B	$\sigma_{rest}^2 + a\sigma_B^2 \times B1$	
$A \times B$	$(a-1)(b-1)$	$M S_{A \times B}$	$\sigma_{rest}^2 + r Q_{A \times B}$	$M S_{A \times B} / E_c$
$E_c(A \times B \times b1)$	$(a-1)(b-1)(r-1)$	E_C	σ_{rest}^2	

表2 条区设计效应均值多重比较的方差与自由度

Table 2 Variances and their degree of freedom of multiple comparisons of means for strip plot design

均值多重比较 Comparison of means	方差 Variance	方差与表1中各均方的关系 Relation of variance to MS	自由度 df
主效应 Main effect			
A	$2(\sigma_{rest}^2 + b\sigma_A^2 \times B1) / (br)$	$2E_A / (br)$	$(a-1)(r-1)$
B	$2(\sigma_{rest}^2 + b\sigma_B^2 \times B1) / (ar)$	$2E_B / (ar)$	$(b-1)(r-1)$
交互效应 Interaction			
同B 异A Different A at same B	$2(\sigma_{rest}^2 + \sigma_A^2 \times B1) / r$	$2[(b-1)E_C + E_A] / (br)$	不可准确确定 Not exactly estimable
同A 异B Different B at same A	$2(\sigma_{rest}^2 + \sigma_B^2 \times B1) / r$	$2[(a-1)E_C + E_B] / (ar)$	不可准确确定 Not exactly estimable
A 和B 都不同 Different B and A	$2(\sigma_{rest}^2 + \sigma_A^2 \times B1 + \sigma_B^2 \times B1) / r$	$2[(ab - a - b)E_C + aE_A + bE_B] / (rab)$	不可准确确定 Not exactly estimable

注: 均值多重比较的标准误为其方差的平方根。

Note: STDERR of multiple comparisons is square root of their variance

综上所述, 条区设计中除 $A \times B$ 交互效应 F 测验的误差为剩余误差 E_C 外, 其余效应的 F 测验以及所有效应均值多重比较(t 测验)的误差均不对应于单一的剩余误差均方 E_C 。其误差相应的自由度也不能仅由一个均方 $M S$ 的自由度来确定。SAS 中的 PROC GLM 和 PROC ANOVA 在进行数据分析时, 系统默认所有测验的误差为剩余误差项。为了得到正确的测验结果, 必须利用语句 TEST H = , E = 和 MEANS E = 对各效应 F 测验和 t 测验相应的误差分别予以定义。

但不可忽视的是, PROC GLM 中只能将方差分析表中的一个均方 $M S$ 定义为效应测验的误差, 对于多误差试验中, 当一个效应的误差为多个均方 $M S$ 的线性组合时, 利用 PROC GLM 既不能对该误差及其自由度按相应均方线性组合自动计算, 也不

能被定义, 因此利用 PROC GLM 不能对模型(1)中区组效应和 $A \times B$ 交互效应的多重比较进行测验。虽然 PROC GLM 会自动给出区组效应 F 测验的结果, 且利用一个语句 LS MEANS A * B / PDIFF 也能得到 $A \times B$ 交互效应多重比较的结果, 但其测验结果是系统自动利用剩余误差及其自由度对该效应进行测验得到的, 其结果是不可靠的。其误差的大小将随方差分量 σ_{rest}^2 , $\sigma_A^2 \times B1$ 和 $\sigma_B^2 \times B1$ 的比例以及试验因素水平 a , b 和 r 比例的不同而变化。

2 PROC MIXED 的原理依据与应用

随着混合线性模型理论的完善以及方差分量估计方法的发展, 利用混合线性模型解决试验统计分析问题的研究和应用近几年受到了广泛的重视^[7]。根据混合线性模型的原理, 不需要依赖方差分析表

计算均方 MS , 而通过直接估算各项随机效应的方差分量, 并进一步估算固定效应均值、效应线性对比值及其标准误和自由度, 就可以进行固定效应统计测验。利用混合模型进行试验分析运算较复杂, 但统计软件 SAS 中 $PROC MIXED$ 的开发可使人们很容易地利用混合线性模型的原理, 分析各种复杂的试验设计, 使分析多误差试验设计与单误差试验设计以及平衡数据与非平衡数据同样容易。 $PROC MIXED$ 的应用也比 $PROC GLM$ 更为简明。

2.1 混合线性模型及其效应测验

混合线性模型一般具有下面的矩阵形式^[8]:

$$y = X\beta + Zu + e \quad (3)$$

式中, y 为试验观测值向量; β 为固定效应向量; u 为随机效应向量; e 为剩余效应向量; X 为固定效应试验设计矩阵; Z 为随机效应试验设计矩阵。

对随机效应和剩余效应的分布分别有: $u \sim N(0, G)$, $e \sim N(0, R)$; 对试验观测值的期望值和协方差分别有: $E(y) = X\beta$, $Var(y) = ZGZ' + R = V$ 。

G 和 R 分别是随机效应方差分量和剩余方差分量的函数。在 $PROC MIXED$ 中利用一个 $MODEL$ 语句定义 X , 而 G 的结构用一个 $RANDOM$ 语句定义。若试验只涉及固定效应, 则表明模型(3)中的 u 为0, 即固定模型是混合模型的特殊形式。

在实际研究中, 方差分量的值未知, 通常用限制性极大似然法(REML)依据样本数据估计方差分量^[9]。与此相应, 试验观测值 y 的协方差阵 V 也成为估计量 $\hat{V}$ 。这时, 固定效应 β 的广义最小二乘估计可由

$$\hat{\beta} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}y \quad (4)$$

给出, 且 $\hat{\beta}$ 估计的误差协方差可由

$$var(\hat{\beta}) = (X'\hat{V}^{-1}X)^{-1} \quad (5)$$

求得。

固定效应的线性对比可根据下面的统计量进行测验^[10]:

对2个固定效应的线性对比:

$$t = \frac{L'\hat{\beta}}{\sqrt{L'(X'\hat{V}^{-1}X)^{-1}L}} \sim t(v) \quad (6)$$

对多个固定效应的线性对比:

$$F = \frac{\hat{\beta}'L(L'(X'\hat{V}^{-1}X)^{-1}L)^{-1}L'\hat{\beta}}{rank(L)} \sim F(rank(L), v) \quad (7)$$

其中, L 是描述固定效应线性对比的行向量或矩阵, $rank(L)$ 表示 L 的秩, v 为测验统计量的自由度(第二自由度)。

2.2 统计量自由度的计算

在 $PROC MIXED$ 中有多个自由度计算方法可供选择, 其中Satterthwaite法应用最为广泛, 可通过选项语句 $DDFM = SATTERTHWAITE$ 予以实现。但这里所谓的Satterthwaite法不是如前所述的利用方差分析有关均方 MS 计算统计量的自由度, 而是利用随机效应方差分量的估计值以及方差分量协方差估计值计算统计量的自由度, 因为混合线性模型中固定效应的方差是根据试验观测值的协方差确定的。该方法是Giesbrecht和Burns^[11]以及Fai和Cornelius^[12]对Satterthwaite方法原理分别在 t 测验和 F 测验中的进一步推广, 其间应用了固定效应误差方差期望值的近似计算及误差方差协方差的近似确定。该方法具体有下面的形式:

$$v = \frac{2 \left(\sum_{p=1}^k M_p \hat{\sigma}_p^2 \right)^2}{\sum_{p=1}^k \sum_{q=1}^k Cov(\hat{\sigma}_p^2, \hat{\sigma}_q^2) M_p M_q} \quad (8)$$

式中, $M_p = L'(X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}Z_p Z_p'\hat{V}^{-1}X(X'\hat{V}^{-1}X)^{-1}L$; $Cov(\hat{\sigma}_p^2, \hat{\sigma}_q^2)$ 为REML法估计方差分量的协方差; k 为模型中所含随机效应(方差分量)的数目; $\hat{\sigma}_p^2$ 为由REML法估计的方差组分值; Z_p 为随机效应试验设计矩阵 Z 中关于第 p 个随机效应的分矩阵。式(8)利用了方差组分的估计值及其估计的方差协方差, 分子表示2倍的固定效应线性对比方差的平方, 分母代表固定效应线性对比方差的协方差。为区别于方差分析中的 t 测验和 F 测验, 依式(6)和(7)进行的 t 测验和 F 测验也被称作Wald- t 测验和Wald- F 测验^[11]。

采用 $PROC MIXED$ 分析多误差试验数据时, 只要将随机误差效应(不包含剩余误差效应)编码在 $RANDOM$ 语句中, $PROC MIXED$ 就可自动给出所有固定效应正确的 F 测验结果, 并且利用其中的语句 $LSMEANS/PDIFF$ 便可自动给出固定效应两两比较的正确测验结果。与 $PROC GLM$ 不同, $PROC MIXED$ 不需要分析者必须熟悉试验模型对应的方差分析表, 不需要根据方差分析表中各项期望均方的构成去定义正确的测验统计量。更为重要的是, $PROC MIXED$ 还能自动给出多误差试验中当效应误差为多个均方线性组合时, $PROC GLM$ 不能给出的正确测验结果。

3 实例数据的分析

下面利用上述方差分析并结合Satterthwaite原

理的方法(手算)及 PROC GLM 和 PROC MIXED 共3种方法,对某一条区设计的试验数据进行分析。该试验A 因素为播种期($a=2$),B 因素为除草方法($b=3$),重复(区组)3次($r=3$),测量数据为小麦小区产量(表3)。

由于两因素条区设计中,A,B 和A×B 3种效应F 测验以及A 和B 效应t 测验的误差项分别对应于方差分析表中的一个均方,所以对这些效应3种方法的测验结果相同。A,B 和A×B 测验大于其F 值的概率($P>F$)分别为0.180,0.020和0.002。因素A 主效应差异不显著,因素B 主效应和A×B 交互效应均达到极显著水平。此外,所有测验 PROC MIXED 的结果与手算的结果一致,因此只将PROC GLM 和PROC MIXED 的结果予以对比说明。区组效应F 测验的误差为多个均方的线性组合,因而其效应测验PROC GLM 和PROC MIXED 的结果差异很大,前者F 值和 $P>F$ 值分别为0.76和0.569,后者对应的值为4.39和0.098。A×B 交互效应多重比较的误差也为多个均方的线性组合,PROC GLM 和PROC MIXED 中的测验结果如表4所示,2种分析方法的结果也表现出了不同程度的差异。例如对A₁B₁与A₁B₃的比较,在PROC GLM 中达到显著水平($P>F=0.039$),而在PROC MIXED 中未达到显

著水平($P>F=0.149$)。由于PROC GLM 只能采用一个均方(剩余效应均方)作为误差,所以所有效应对比的标准误和自由度相同。PROC MIXED 按照观测值的协方差与固定效应线性对比计算误差,并基于推广了的Satterthwaite 法确定其自由度,因而计算得到的各类效应对比的误差和自由度相应于误差期望值的不同而变化。由本试验数据计算的3个方差分量 $\sigma^2_{A \times B1}$, $\sigma^2_{B \times B1}$ 和 σ^2_{rest} 的值分别为0.341,7.672和3.416, $\sigma^2_{A \times B1}$ 的值相对于 σ^2_{rest} 值较小,所以B 相同A 不同水平间(A₁B₁-A₂B₁和A₁B₃-A₂B₃)t 测验PROC GLM 与PROC MIXED 的结果差异很小; $\sigma^2_{B \times B1}$ 的值相对于 σ^2_{rest} 值较大,因而A 相同B 不同水平间(A₁B₁-A₁B₂,A₂B₁-A₂B₃和A₁B₁-A₁B₃)以及A 与B 水平都不同(A₁B₁-A₂B₂和A₁B₃-A₂B₂)时,t 测验2种方法的结果差异很大。

表3 文中分析所用小麦试验产量

Table 3 Data of wheat yield used in the text

区组 Block	A 1			A 2		
	B 1	B 2	B 3	B 1	B 2	B 3
1	73.4	60.8	82.3	78.0	75.3	75.1
2	84.0	64.2	81.7	85.5	73.4	75.0
3	78.6	62.5	85.7	76.0	76.2	76.8

表4 PROC GLM 和PROC MIXED 对A×B 交互效应进行多重比较(t 测验)的部分结果

Table 4 Multiple t-test for A×B interaction in PROC GLM and PROC MIXED (selected results)

效应比较 Effect comparisons				PROC GLM				PROC MIXED			
A _i	B _j	A _{i'}	B _{j'}	标准误 Stderr	df	t 值 t value	P> t *	标准误 Stderr	df	t 值 t value	P> t *
1	1	2	1	1.509	4	-0.77	0.482	1.582	5.90	-0.74	0.489
1	3	2	3	1.509	4	5.04	0.007	1.582	5.90	4.80	0.004
1	1	1	2	1.509	4	10.71	0.000	2.719	5.41	5.95	0.002
2	1	2	3	1.509	4	2.78	0.050	2.719	5.41	1.54	0.178
1	1	1	3	1.509	4	-3.03	0.039	2.719	5.41	-1.68	0.149
1	1	2	2	1.509	4	2.45	0.070	2.760	5.64	1.34	0.232
1	3	2	2	1.509	4	5.48	0.005	2.760	5.64	2.99	0.026

注: $P>|t|$ 表示大于t 值绝对值的概率(显著水平)。

Note: $P>|t|$ is the probability level

4 结论与建议

试验模型中有多个误差成分时,如利用PROC GLM 进行固定效应测验,则必须对固定效应相应的误差项给予明确定义方可得到正确的测验结果,这要求分析者具备足够的方差分析知识。此外,在有多个误差来源的试验模型中,可能会存在效应比较的误差为多个误差均方线性组合的情况,这时PROC GLM 不能对该效应进行正确的统计测验。

本文仅涉及了条区设计,且A 和B 2个试验因素都为固定效应的情况,此时其主效应和交互效应F 测验的误差均对应于方差分析表中的一个均方,所以利用PROC GLM 分析时,两因素主效应及其交互效应的F 测验不存在问题;如果F 测验交互效应不显著,这时交互效应的多重比较(t 测验)也就没有必要进行,那么整个分析结果也不存在问题。但是如果交互效应F 测验显著,需要进一步进行交互作用的多重比较,这时PROC GLM 就不能得到可

靠的多重比较结果。此外,如果有一个因素,例如B因素为随机因素,这时A因素 F 测验对应的误差就变为几个均方的线性组合,裂区设计也是如此^[13],那么PROC GLM对A因素也不能进行正确的 F 测验。一个弥补的办法就是利用PROC GLM计算得到的效应估计值和各项均方值,按照效应对比误差的线性关系式手工计算相应的误差值,并结合Satterthwaite法计算自由度,然后取得正确的测验结果。但如果试验数据又为非平数据,那么手工计算就可能非常复杂或者根本不可能。

PROC MIXED利用了混和线性模型的原理以及广义的Satterthwaite确定自由度的方法,既能得到PROC GLM给出的正确测验结果(误差项为单个均方时),又能得到PROC GLM不能直接给出的正确测验结果(效应误差为多个均方的线型组合时),而且应用者不必去熟悉方差分析表的原理和内容,进而定义效应测验对应的误差,既适用于平衡数据,又适用于非平衡数据,使用简单,结果可靠。鉴于此,建议利用PROC MIXED分析多误差试验设计的数据。

[参考文献]

- [1] Piepho H P, Spilke J. Anmerkungen zur analyse balancierter gemischter Modelle mit der SAS-Procedure MIXED[J]. Zeitschrift fuer Agrarstatistik, 1999, 2: 39- 46
- [2] 南京农业大学. 田间试验和统计方法[M]. 北京: 中国农业出版社, 1993 184- 190
- [3] Harville D A. A lternative fomulation and procedure for the two-way mixed model[J]. Biometrics, 1978, 34: 441- 453
- [4] Milliken G A, Wolliams E R. Approximating degree of freedom for standard errors in mixed linear models[A]. Proceedings of the Statistical Computing Section[C]. New Orleans: American Statistical Association, 1985 50- 59
- [5] Satterthwaite F E. Synthesis of variance[J]. Psychometrika, 1941, 6: 309- 316
- [6] Satterthwaite F E. An approximation distribution of estimates of variance components[J]. Biometrics Bulletin, 1949, 2: 110- 114
- [7] McCulloch C E, Searle S R. Generalized, linear and mixed models[M]. New York: John Wiley and Sons, 2001
- [8] Henderson C R. Statistical method in animal improvement[A]. Advances in statistical methods for genetic improvement of livestock[C]. New York: Springer-Verlag, 1990
- [9] Searle S R, Casella G, McCulloch C E. Variance components[M]. New York: John Wiley and Sons, 1992
- [10] Verbeke G, Molenberghs G. Linear mixed models in practice[M]. New York: Springer-Verlag, 1997
- [11] Giesbrecht F G, Burns J C. Two-stage analysis based on a mixed model: large- sample asymptotic theory and small-sample simulation results[J]. Biometrics, 1985, 41: 477- 486
- [12] Fai A H T, Cornelius P L. Approximate F -tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments[J]. J Statist Comput Simul, 1996, 54: 363- 378
- [13] 袁志发, 周静芊. 试验设计与分析[M]. 北京: 高等教育出版社, 2000 248- 263

Application of PROC MIXED in analysis of the trials with more than one error

HU Xi-yuan

(College of Agronomy, Northwest A & F University, Yangling, Shaanxi 712100, China)

Abstract By applying the strip-plot design as an example, the characteristics of trial analysis with more than one error and the limitations of analyzing this type of experiments using PROC ANOVA and PROC GLM in software SAS were discussed. And the principle and the advantage of analyzing trial data with PROC MIXED were introduced and discussed. Then suggestions were given by using PROC MIXED to analyze the trials with more than one error.

Key words: trial analysis; PROC GLM; PROC MIXED