一九八六年　西南农业大学学报　№. 4
第四期　Journal of Southwest Agricultural University　Nov. 1986

重复的二次回归正交旋转设计试验数学模型的选择

农学系　罗泽伟

【提要】由赤池统计模型推断的信息量准则（AIC）构造出作为二次回归正交旋转设计试验指导的回归模型的AIC统计量，其形式为：

$$AIC=N\ln\delta^2+\frac{1}{\delta^2}(y-x\hat{\beta})^T(y-x\hat{\beta})+2K$$

利用以上准则评判具有重复的二次回归正交旋转设计试验多个数学模型的选择。

多因子二次回归正交旋转设计是将试验设计、数学模型的建立和模型的稳定性与精确性综合考虑的多因子试验设计[1]。近年来，我国农业科技工作者又将其应用于农作物栽培规范化农艺措施的研究中，已获得很大的社会和经济效益。对于大面积大幅度提高作物产量起到巨大的推动作用。

二次回归设计(具有正交性或旋转性是将试验点均匀地分别在P维因子空间中半径分别为$\rho=0$、$\rho=+1$、$\rho=+\gamma$的三个球面上的，因而将试验点分为三类：分布在原点的零水平试验点m_0个；分布在单位球面上的试验点m_c个；分布在星号位球面上的试验点m_r个。总的试验次数为$N=m_0+m_c+m_r$。试验不设置重复。

农业田间试验由于外因素的干扰，误差较大。根据同一试验不同重复(地点、时间)所获得的数据建立的数学模型往往有较大的差异。因此，为了提高试验的灵敏度、降低试验误差对模型的干扰、确保所得的概括大范围作物栽培生产过程的数学模型的代表性，试验需要重复。本文应用赤池信息量准则构造出二次回归正交旋转设计不同重复试验数学模型的选择标准。

一、基本原理

P因子二次回归正交旋转设计试验指导的数学模型其一般形式为：

$$Y=\beta_0+\sum_{i=1}^{p}\beta_iZ_i+\sum_{1\le i<j\le p}\beta_{ij}Z_iZ_j+\sum_{j=1}^{p}\beta_{jj}Z_j^2 \quad \cdots\cdots ①$$

若记，

$$Z=(1\ \ Z_1\cdots Z_p\ \ Z_1Z_2\cdots Z_{p-1}Z_p\ \ Z_1^2\cdots Z_p^2)^T$$

$$\beta=\beta_0\beta_1\cdots\beta_p\ \ \beta_{12}\cdots\beta_{p-1p}\ \ \beta_{11}\cdots\beta_{pp})^T$$

则①式可以写成向量形式：$Y=Z^T\beta$ ……②

试验设计的结构矩阵及目标观察值矩阵见表1。

表1　P因子二次回归正交旋转设计的结构矩阵

	试验号	z_0	z_1	$z_2 \cdots z_p$	z_1z_2	$z_1z_3 \cdots z_{p-1}$	$z_pz_1^2$	$z_2^2 \cdots z_p^2$	重复 I	重复 II	…	重复 r
m_c	1	1	1	1…1	1	1 … 1	1^*	$1^* \cdots 1^*$	y_{11}	y_{12}	…	y_{1r}
	2	1	1	1…−1	1	1 … 1	1^*	$1^* \cdots 1^*$	y_{21}	y_{22}	…	y_{2r}
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
	2^p	1	−1	−1…−1	1	1 … 1	1^*	$1^* \cdots 1^*$				
m_r	2^p+1	1	r	0…0	0	0 … 0	r^{*2}	$0^* \cdots *$				
	2^p+2	1	−r	0…0	0	0 … 0	r^{*2}	$0^* \cdots 0^*$				
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮				
	2^p+2p	1	0	0…−r	0	0 … 0	0^*	$0^* \cdots 0^*$				
m_0	2^p+2p+1	1	0	0…0	0	0 … 0	0^*	$0^* \cdots 0^*$				
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮				
	N	1	0	0…0	0	0 … 0	0^*	$0^* \cdots 0^*$	y_{N1}	y_{N2}	…	y_{Nr}

其中：①$N=m_c+m_r+m_0$；②r是设计的臂长参数；③"*"表示经中心化处理的数据。

$$X=\begin{pmatrix} 1 & 1 \cdots 1 & 1 \cdots 1 & 1^* \cdots 1^* \\ \vdots & \vdots & \vdots & \vdots \\ 1 & r \cdots 0 & 0 \cdots 0 & r^{*2} \cdots 0^* \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 \cdots 0 & 0 \cdots 0 & 0^* \cdots 0^* \end{pmatrix} \qquad \overline{Y}=\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix}$$

则，②中未知参数向量β的极大似然估计为：

$$\hat{\beta}=(X^TX)^{-1}X^TY \quad \cdots\cdots ③$$

于是模型②的理论预测值向量为：

$$\hat{Y}=\begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_N \end{pmatrix}=X\hat{\beta}$$

假定随机向量：$\varepsilon \hat{=} Y-\hat{Y}=X\beta-X\hat{\beta} \sim N(0, \sigma^2 I_N)$

根据赤池信息准则，模型的信息量推断统计量是：

$$AIC=-2\ln L(\hat{\beta},\ \sigma^2;\ Z)+2K \quad \cdots\cdots④$$

其中，$L(\hat{\beta},\ \sigma^2;\ Z)$是模型的极大似然函数，K是模型中的有效参数个数。

在此，由于 $\varepsilon=Y-X\hat{\beta}\sim N(0,\ \sigma^2 I_N)$

则，$$L(\hat{\beta},\ \sigma^2;\ Z)=\frac{N}{\prod_{i=1}}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(\overline{Y}-X\hat{\beta})^T(\overline{Y}-X\hat{\beta})\right\}$$

$$\ln L(\hat{\beta},\ \sigma^2;\ Z)=\sum_{i=1}^{N}\ln\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(Y-X\hat{\beta})^T(Y-X\hat{\beta})\right\}$$

$$=-\frac{N}{2}\ln(2\pi)-\left\{\frac{N}{2}\ln\sigma^2+\frac{1}{2\sigma^2}(Y-X\hat{\beta})^T(-Y-X\hat{\beta})\right\}$$

在上式中，由于$-\frac{N}{2}\ln(2\pi)$是一个常数因子，可将其略去得到：

$$\ln L(\hat{\beta},\ \sigma^2;\ Z)=-\frac{N}{2}\ln\sigma^2-\frac{1}{2\sigma^2}(Y-X\hat{\beta})^T(Y-X\hat{\beta})$$

将上式代入④，得到：

$$AIC=N\ln\sigma^2+\frac{1}{\sigma^2}(Y-X\hat{\beta})^T(Y-X\hat{\beta})+2K \quad \cdots\cdots⑤$$

⑤式的含意是：它可以作为某一个统计模型，其建立依据的数据资料所包含的信息量以及模型中的估计参数对这些信息的概括能力和稳定性能的模型推断综合指标。利用AIC对多个模型进行选择时，是遵循最小值原理，即是相应的AIC为最小的模型被认为是可取的。

在表1中，利用不同重复的目标观察值为依据可以获得不同的数学模型。例如第t个重复相应的模型为：

$$\hat{Y}t=b_0^{(t)}+\sum_{i=1}^{p}b_i^{(t)}Z_i+\sum_{1\le i\le j\le p}b_{ij}^{(t)}Z_iZ_j+\sum_{j=1}^{p}b_{jj}^{(t)}Z_j^2$$

相应的模型推断统计量为：AIC_t　　$(t=1,\ 2,\ \cdots,\ r)$

若记，$$AIC_{i_0}^{*}=\min_{1\le i\le r}\{AIC_i\}$$

则AIC_{i_0}对应的模型：

$$\hat{Y}_{i_0}=b^{(i_0)}+\sum_{i=1}^{\rho}b_i^{(i_0)}Z_i+\sum_{1\le i\le j\le\rho}b_{ij}^{(i_0)}Z_iZ_j+\sum_{j=1}^{\rho}b_{jj}^{(i_0)}Z_j^2$$

是在上述意义下的最佳模型。

表2　四次重复试验各回归模型的回归系数表

重　复	1	2	3	4
b_0	121.99	120.77	120.83	120.69
b_1	-4.42	-4.04	-3.29	-3.29
b_2	0.50	0.71	1.38	1.21
b_3	-0.33	-0.13	-0.29	-0.38
b_4	0.50	-0.13	0.38	0.21
b_5	0.42	0.04	0.04	0.54
b_{12}	0.27	1.22	0.63	0.84
b_{13}	0.02	-0.66	-0.36	-0.41
b_{14}	0.27	-0.28	-0.48	-0.28
b_{15}	0.02	-0.28	-0.49	-0.28
b_{23}	-0.11	-0.28	-0.61	-0.78
b_{24}	0.63	0.31	0.19	0.56
b_{25}	-0.38	-0.06	0.06	0.31
b_{34}	-0.63	-0.06	-0.19	-0.06
b_{35}	-0.63	-0.06	0.19	-0.06
b_{45}	0.50	-0.06	0.31	-0.19
b_{11}	0.75	-0.06	0.31	0.19
b_{22}	0.50	-0.06	0.19	-0.06
b_{33}	-0.75	0.06	0.44	-0.06
b_{44}	-0.50	0.06	-0.19	-0.06
b_{55}	-0.25	0.06	-0.19	0.06

二、应用实例

本段引用“重庆市杂交水稻规范化栽培技术研究”课题的五因子：播种期（Z_1）、移栽叶龄（Z_2）、栽植密度（Z_3）、施氮水平（Z_4）、施磷钾水平（Z_5），于实施的二次回归正交旋转设计重复4次试验的播种—齐穗天数数据资料为例，引证本文介绍的方法。

根据Robert R. Sokal建立的各重复相应二次回归模型的回归系数列入表2；显著性检验见表3：

表3　各次重复回归模型的显著性检验

重复 / 变因	1	2	3	4
SS总	563.64	538.75	837.89	422.75
SS回	544.06	475.91	365.10	364.23
SS误	2.90	4.50	2.40	2.50
SS失	16.67	58.34	70.39	52.93
SS离	19.57	62.84	72.78	55.43
F　值	11.26	2.94	1.91	2.54
显著性	0.01	0.05	ns	0.05

由表3可知，除第3重复外其余各重复试验均达到0.05以上的显著性水平。为了进一步判断模型的稳定性，在AIC准则意义下，即模型中的有效（显著）参数个数，得到各重复相应的回归模型中各偏回归系数的显著性分析（表4）。

表4表明，第Ⅰ重复相应的模型中显著（$P<0.05$）的参数个数为16与另外三个模型相比达到最大；第Ⅱ重复相应的模型中显著（$P<0.05$）的参数个数为7与另外三个模型比较为最小。由于模型的稳定性是随模型中所包含的参数增加而下降的。因此，稳定性指标以模型Ⅰ为最差，模型Ⅱ最优。但是，从表3的分析结果表明，由于模型Ⅰ的显著水平最高，因而拟合性最好。这样看来模型的拟合性和稳定性之间有时会出现矛盾。赤池弘次认为，构造的模型推断准则是综合了矛盾两方面协调作用的统计量。

根据⑤式，以$N=36$次试验，$\hat{\sigma}^2$为试验误差的方差估计量，$(Y-X\hat{\beta})^T(Y-X\hat{\beta})$为离回归变异平方和，K是模型中所含的有效参数个数。求出各试验重复模型的AIC统计量：

表4　各重复回归模型中偏回归系数的显著性分析

变因 \ 效应项	一次项					交互项										二次项				
	X_1	X_2	X_3	X_4	X_5	X_1X_2	X_1X_3	X_1X_4	X_1X_5	X_2X_3	X_2X_4	X_2X_5	X_3X_4	X_3X_5	X_4X_5	X_1^2	X_2^2	X_3^2	X_4^2	X_5^2
Ⅰ重复 DF	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
MS	468.17	6	2.67	6	4.17	2.35	0.01	2.35	0.01	0.01	6.25	2.25	9.25	6.25	4	9	4	9	4	1
F	1452.93	18.62	8.28	18.62	12.93	7.28	0.04	7.28	0.04	0.04	19.39	6.98	19.39	19.39	12.41	27.93	12.41	27.93	14.41	3.10
显著	0.01	0.01	0.05	0.01	0.01	0.05	ns	0.05	ns	ns	0.01	0.05	0.01	0.01	0.01	0.01	0.01	0.01	0.01	ns
Ⅱ重复 DF	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
MS	784.04	12.01	0.38	0.38	0.04	47.53	13.78	2.53	2.53	2.53	1.56	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
F	784.08	24.08	0.75	0.75	0.08	95.36	27.56	5.06	5.06	5.06	3.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13
显著	0.01	0.01	ns	ns	ns	0.01	0.01	0.05	0.05	0.05	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
Ⅲ重复 DF	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
MS	260.04	46.38	2.04	3.38	0.04	12.92	4.25	7.67	7.67	12.09	0.56	0.06	0.56	0.56	1.57	1.57	0.56	3.06	0.56	0.56
F	975.15	170.15	7.66	12.66	0.19	48.45	15.95	28.76	28.76	45.33	2.11	0.23	2.11	2.11	5.85	5.85	2.11	11.48	2.11	2.11
显著	0.01	0.01	0.05	0.01	ns	0.01	0.01	0.01	0.01	0.01	ns	ns	ns	ns	0.05	0.05	ns	0.01	ns	ns
Ⅳ重复 DF	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
MS	260.04	35.04	3.38	1.04	7.04	22.78	5.28	2.53	2.53	19.53	5.06	1.56	0.06	0.06	0.56	0.56	0.06	0.06	0.06	0.06
F	936.15	126.15	12.15	3.75	25.35	82.01	19.01	9.11	9.11	70.31	18.23	5.63	0.23	0.23	2.03	2.03	0.23	0.23	0.23	0.23
显著	0.01	0.01	0.01	ns	0.01	0.01	0.01	0.05	0.05	0.01	0.01	0.05	ns	ns	ns	ns	ns	ns	ns	ns

AIC_1：104.48　　AIC_2：231.41　　AIC_3：560.89　　AIC_4：365.98

于是可以推断，在模型的拟合性和稳定性下，四个模型的优良性顺序是：Ⅰ、Ⅱ、Ⅳ、Ⅲ。

在本例中，模型的拟合性是最佳模式推断的主要因素。事实上，尽管模型Ⅰ的稳定性最差，但由于它的拟合性比其它模型强得多（显著性水平的数量级差别），因而弥补了稳定性方面的不足，最终被选为最佳模型。第Ⅱ模型的AIC估计量与第Ⅳ模型相应的AIC值小134.57。这反映出了这两个几乎相同拟合性的模型在稳定性方面表现出的巨大差别，即在综合试验资料提供的信息量能力方面，模型Ⅱ比之模型Ⅳ优良得多。因此，在只有Ⅱ、Ⅳ两模型供选择时，按AIC准则，应该毫不迟疑地选择前者作为生产过程的预测和控制的数学模型。笔者认为，从AIC准则和统计模型的优良性推断，能尽可能将模型的拟合性集中到少数几个参数上的模型是好的模型。

利用二次回归正交旋转设计试验建立刻划某一生产过程的数学模型，其目的的还主要在于研讨该过程中涉及的主要动态因子的数量关系，进而制订生产过程的最佳决策方案。因此，数学模型的代表性是必须的基础，而拟合性及稳定性则又是统计模型代表性的主要内容。可见本文构造出的二次回归模型的优良性推断准则，其实际意义是显然的。

参　考　文　献

①茆诗松等：回归分析及其试验设计，华东师大出版社，1981

②朱伟勇：回归设计及其应用，数学的实践与认识（3），1978

③朱伟勇：回归设计及其应用，数学的实践与认识（4），1978

④刘璋温：选择回归模型的几个准则，数学的实践与认识（4），1983

⑤刘璋温：赤池信息量准则AIC及其意义，数学的实践与认识（3），1980

⑥重庆市《杂交中稻高产栽培技术规范研究》协作组：杂交中稻高产栽培技术规范研究试验、示范总结，西农科技（3），1986

⑦Robert, R. Sokal: Biometry, W. H. Freeman and Compary, 1969

CHOICE OF MATHEMATICAL MODEL FROM REGRESSIONAL ORTHOGONAL AND VOTATIONAL DESIGN EXPERIMENT AT TWO EXPONENTS (ROVDETE) WITH REPLICATION

Luo Zewei

(Department of Agronomy, Southwest Agricultural University)

Abstract

The statistics, a H. Akaike information criterion dealt with ROVDETE was constructed according to information criterion concerned about statistical model inference presented by H. Akaike. It had the following general form:

$$AIC=N*\ln\sigma^2+(Y-X*\hat{\beta})^T*(Y-X*\hat{\beta})/\sigma^2+2*K$$

The formula above could be used to inference the choice of multi—mathematical model from the replicated ROVDETE.