偏最小二乘回归分析在均匀设计试验建模分析中的应用

唐启义

(浙江大学农业与生物技术学院　杭州)

摘要　本文分析了目前应用一般的最小二乘法建立均匀试验数据的二次多项式回归模型时存在的局限性，提出了应用偏最小二乘法(Partial least-square，PLS)建立二次多项式回归模型的技术，并已在作者开发的统计分析软件(DPS 数据处理系统)中实现。然后以一实例对 PLS 的回归建模过程进行了介绍。作者认为，PLS 回归分析建模技术将为均匀设计的更广泛应用提供有力的技术支持。

关键词　偏最小二乘法，均匀设计，回归分析，模型优化

引言

回归分析是均匀设计数据分析的主要手段。由于均匀设计的出发点是建立多因素寻优模型，这样，如考虑多因素互作、模型最优化的的实际需要，最基本的要求是根据均匀设计试验结果建立二次多项式回归模型。若试验设计有 m 个因素 $x_1,\cdots,x_m$，当观察指标为 y 时，其二次多项式回归模型为

$$y=\beta_0+\sum_{i=1}^{m}\beta_i x_i+\sum_{i=1}^{m}\beta_{ii}x_i^2+\sum_{i<j}\beta_{ij}x_i x_j+\varepsilon$$

其中β_0，β_I，β_{ii}和β_{ij}为回归系数，ε 为随机误差。从上述回归模型可以看到，除了常数项β_0 以外，方程有 $m(m+3)/2$ 项，若使回归系数的估计有可能，必要条件为试验次数 $n>1+m(m+3)/2$。当 m 较大时，通常不能满足这个必要条件。目前通常的做法是采用逐步回归分析技术，从二次多项式方程中选择方差贡献显着的因素或因素组合，删除不显着（重要）的因素或因素组合，建立含部分变量的回归方程模型。

但是，从实际操作、应用来看，有几个问题：一是分析时，多数自变量是组合变量，它们之间存在有严重的多重共线性，这会使得分析结果很不稳定，以致有时，某个因素是否选入对回归方程产生很大的影响，使建模者左右为难；二是选中的自变量，有时与我们所希望的有较大的出入，从专业知识方面认为是重要的变量往往落选，特别是有时单相关非常显着的变量落选，使我们很难信服地接受这样的“最优”回归模型；三是所建立的回归方程模型，有的因素的回归系数符号反常，这与专业背景不符合；四是在配方均匀设计试验、并考查外界影响因素

时，配方成份是不能随意去掉的；最后，我们有时考察试验结果是多个指标，对多个目标变量同时建模分析，这是一般的最小二乘回归分析方法不能解决的。从上述这 5 个问题可以看出，传统的基于最小二乘的多元线性回归、逐步回归分析方法不能完全适应均匀设计数据建模过程的需要。

偏最小二乘法(partial least squares)回归分析方法，这一从应用领域提出的一种新的多元数据分析技术在近 10 多年以来得到了迅速地发展。偏最小二乘法可以有效地克服目前回归建模的许多实际问题，如上面提到的样本容量小于变量个数时进行回归建模，以及多个因变量对多自变量的同时回归分析等一般最小二乘回归分析方法无法解决的问题。

1. 基本理论与算法简介

偏最小二乘回归分析，最初是研究多解释变量和多个反应变量的定量关系，即在解释变量空间和反应变量空间分别寻找某些线性组合（潜变量），并使得两个变量空间的协方差最大。如用 $\boldsymbol{X}_{n\times m}$ 表示解释变量，用 $\boldsymbol{Y}_{n\times k}$ 表示反应变量，这里 n 是样本个数。PLS 的目的是将数据集投影到一系列的潜变量 $\boldsymbol{t}_j$ 和 $\boldsymbol{u}_j$（j=1,2,…..A），这里 A 是潜变量的个数。然后在 $\boldsymbol{t}_j$ 和 $\boldsymbol{u}_j$ 之间建立回归方程

$$\boldsymbol{u}_j=b_j\boldsymbol{t}_j+\mathbf{e}_j$$

这里的 $\boldsymbol{e}_j$ 是误差向量，b_j 是未知参数。且 b_j 可通过公式 $\hat{b}_j=\left(t_j^Tt_j\right)^{-1}t_j^Tu_j$ 进行估计。

潜变量可通过公式 $\boldsymbol{t}_j=\boldsymbol{X}_j\boldsymbol{w}_j$ 和 $\boldsymbol{u}_j=\boldsymbol{Y}_j\boldsymbol{q}_j$ 计算得到。这里变量 $\boldsymbol{w}_j$ 和 $\boldsymbol{q}_j$ 是使得潜变量 $\boldsymbol{t}_j$ 和 $\boldsymbol{u}_j$ 的协方差最大，亦即使潜变量 $\boldsymbol{t}_j$ 和 $\boldsymbol{u}_j$ 相关程度达最大时的权重系数；

$$X_{j+1}=X_j-t_jp_j^{\mathrm{T}},X_j=X,\quad p_j=X_j^{\mathrm{T}}t_j/(t_j^{\mathrm{T}}t_j)$$

$$Y_{j+1}=Y_j-b_jt_jq_j^{\mathrm{T}},\quad Y_j=Y,\quad p_j=X_j^{\mathrm{T}}t_j/(t_j^{\mathrm{T}}t_j)$$

设 $\hat{u}_j=\hat{b}_jt_j$ 是 $\boldsymbol{u}_j$ 的预报值，这时矩阵 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 可以分解成如下外积形式：

$$X=\sum_{j=1}^{A}t_jp_j^{\mathrm{T}}+E,\qquad Y=\sum_{j=1}^{A}\hat{u}_jq_j^{\mathrm{T}}+F$$

这里 $\boldsymbol{E}$ 和 $\boldsymbol{F}$ 是提取 A 对潜变量后矩阵 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的残差。

在偏最小二乘回归分析过程中，每对潜变量 $\boldsymbol{t}_j$ 和 $\boldsymbol{u}_j$（j=1,2,…..A）在迭代过程中依次被提取，然后计算提取后的残差，并对每一步的残差再继续进行分析，直至根据某种准则确定提取潜变量的对数(A)。

确定要提取的潜变量对数一般是应用预测残差平方和 PRESS(Predicted Residual Sum of Squares)，即在每一步分别计算去掉 1 个样本点后反应变量预测估计值和实际观测值的残差平方和：

$$PRESS_{(j)} = \sum_{k}^{l}\sum_{i}^{n}\left(y_{ik} - \hat{y}_{k(j)(-i)}\right)^2$$

如果 $PRESS_{(j)}$-$PRESS_{(j-1)}$小于预定精度，那么迭代过程结束，否则继续提取潜变量，进行迭代计算。但在实际工作中，可以根据 *PRESS* 的变化，并结合拟合残差平方和的变化趋势进行判断，人为指定提取潜变量的个数。作者认为，应用偏最小二乘建立二次多项式回归模型，提取潜变量个数最多不要超过试验处理的因子个数。

上述偏最小二乘回归分析技术，作者已用 Pascal 程序语言实现，并辅助以图形方式的工作界面，让使用者决定提取潜变量的个数。在建立模型后，系统立即对模型进行优化（求最大值或最小值）。这些功能作为一个统计分析模块收录在作者开发的通用统计分析软件包“ DPS 数据处理系统 ”之中，其演示版本可从网站 http://www.chinadps.net 下载试用。

2 应用实例

张承恩(http://ust40.html.533.net)在研究 VD_3 合成过程中，对其中的一步光化学反应，采用均匀设计技术设计了一套试验 4 个处理因素、7 个处理水平的试验方案，做了 7 批试验，考察了 2 个独立指标和一个复合指标，其试验处理及结果如表 1。

表 1　均匀设计试验数据

投料量	某溶剂量	反应时间	反应温度	转化率	精制率	收率
x_1	x_2	x_3	x_4	y_1	y_2	y_3
30	405	1.5	47.5	76.7	52.2	40.0
40	435	3.0	45.0	84.3	53.4	45.0
50	465	1.0	42.5	65.6	38.7	25.4
60	390	2.5	40.0	69.3	37.1	25.7
70	420	0.5	37.5	38.6	46.3	20.0
80	450	2.0	35.0	58.1	34.4	20.0
90	480	3.5	50.0	59.3	37.3	22.1

在该试验中，有 4 个处理因素，如果建立完整的二次多项式回归方程，需要 15 个处理组合，但这里只有 7 个处理组合，因此只能应用逐步回归分析法，选出较“ 重要 ”的因素或变量组合建立回归方程。对这 3 个产出指标，也只能分别建立 3 个回归方程。如果应用逐步回归分析方法进行建模，就有可能因引入/剔除变量的 F_x 临界值不同，不同的建模人员建立的方程会有很大的差异，并给模型的整体优化，寻求最好的工艺条件等实际应用带来困难。

根据该试验结果，作者应用偏最小二乘回归分析方法，借助于作者编制的偏最小二乘回归分析程序进行分析。分析时参考 *PRESS* 统计量和误差统计量的下降趋

势（图１），选取３个潜变量（组分）来建立二次多项式回归模型。

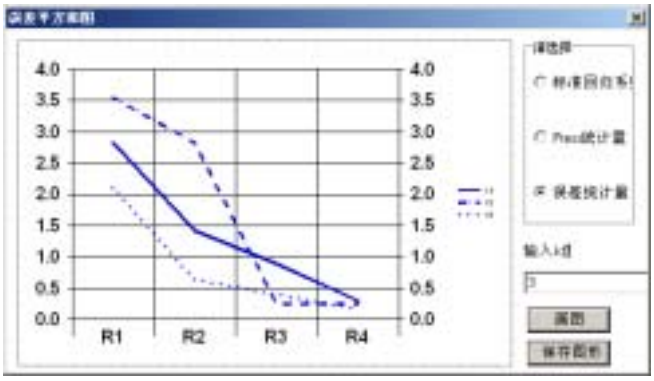

图１　偏最小二乘回归误差统计量下将趋势图

选择潜变量个数，一般认为是根据偏最小二乘分析程序的提示(图 2)进行，从图 2 可以看出，当潜变量个数为 3 时，x_2 和 x_3 的误差已接近为零，且以后变化趋于平缓。不过，根据作者看法，须以各个效应的标准回归系数作为参考，即要与实际的专业背景相吻合。

在选择潜变量个数后，我们再确定模型优化的一些条件（图 2），如决定自变量是否受配方条件的限制，即哪些自变量之和为 1；以及因变量的优化方向（是求极大值还是求极小值）。对有 l 个因变量的系统优化，其目标函数的定义为 $\sum_{i=1}^{l} d \cdot \left(\frac{\hat{y}_i - \bar{y}_i}{SD_i} \right) \cdot 100$，当某目标函数是求极大值时，$d$=1，否则 d=–1。式中 $\hat{y}_i$、$\bar{y}_i$ 和 SD_i 分别是第 i 个因变量的理论值、均值和标准差。

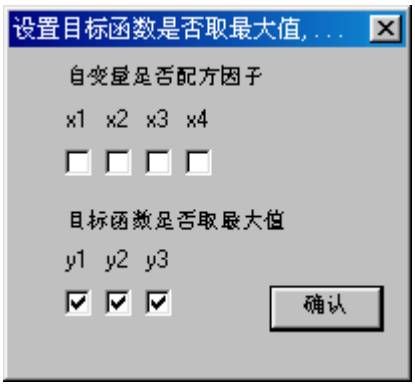

图 2 偏最小二乘回归模型优化

选择有关参数并确认后，我们可得到各个自变量对各个因变量作用的主效应的标准回归系数如表 2（各个自变量的交互作用项的标准回归系数在此略去）。从表 2 可以看出，各个自变量对 3 个因变量的影响是相同的，但自变量 x_1 和 x_2 对因变量的作用为负效应，而 x_3 和 x_4 对因变量的作用为正效应。

表２　各个自变量对各个因变量作用的标准回归系数

考察指标	x1	x2	x3	X4
y1	-0.32017	-0.05775	0.25902	0.29784
y2	-0.36651	-0.13511	0.01560	0.29441
y3	-0.40514	-0.11299	0.16486	0.35029

最后，我们根据偏最小二乘回归分析，同时考虑 3 个因变量的优化，得到如下二次多项式回归模型：

y_1=-220.7351-2.0870x_1-0.1579x_2+54.4213x_3+15.1891x_4+0.007466x_1^2+0.000508x_2^2-3.1747x_3^2-0.1055x_4^2+0.002536x_1x_2+0.01375x_1x_3-0.003724x_1x_4-0.03386x_2x_3-0.009298x_2x_4-0.5714x_3x_4

y_2=-629.4643-0.4055x_1+2.6433x_2-33.2815x_3+6.8048x_4+0.004211x_1^2-0.003051x_2^2+2.01072x_3^2-0.05955x_4^2+0.000400x_1x_2-0.09135x_1x_3-0.005254x_1x_4+0.048726x_2x_3-0.003350x_2x_4+0.2267x_3x_4

y_3=-593.0145-1.1637x_1+1.9208x_2-2.8398x_3+11.3205x_4+0.006231x_1^2-0.002085x_2^2+0.2154x_3^2-0.08792x_4^2+0.001344x_1x_2-0.06298x_1x_3-0.005448x_1x_4+0.02289x_2x_3-0.006306x_2x_4-0.06272x_3x_4

这 3 个二次多项式回归模型的拟合效果，可从误差平方和看出（表 3）。表 3 中显示出提取不同潜变量个数时数据标准化后模型误差平方和和 Press 统计量下降情况，并可得到相应组分时的模型拟合的决定系数 R^2。从决定系数可以看出，提取 3 个组分（潜变量）时，各个回归模型的拟合程度都较好。

表 3 数据标准化后模型误差平方和及决定系数

潜变量个数	误差平方和			决定系数 R^2			Press 统计量		
	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
1	2.81833	3.54253	2.08920	0.53028	0.40958	0.65180	4.26100	5.40944	4.27302
2	1.42179	2.82031	0.62936	0.76303	0.52995	0.89511	3.55193	5.29229	3.08667
3	0.89157	0.24953	0.38381	0.85140	0.95841	0.93603	4.03162	4.99766	2.82919

根据图 2 设定的优化条件对各个模型优化后，得到各个自变量的优化值分别为：x_1 等于 30.0000，x_2 等于 418.2546，x_3 等于 3.5000，x_4 等于 47.2173，综合指标的最优目标函数为 144.97。这时，各个因变量的最优目标函数值 y_1=80.17, y_2=61.30, y_3=48.69。此优化结果作者已反馈给张承恩先生，经张先生确认，其优化方向与原回归分析基本一致，具有合理性，可用。

3.讨论

偏最小二乘回归(Partial least－squares regression)是一种新型的多元统计数据分析方法，它于 1983 年由伍德(S．Wold)和阿巴诺(C．Albano)等人首次提出。近

十几年来，它在理论及应用方面都得到了迅速发展。偏最小二乘回归由于集多元线性回归分析、典型相关分析和主成分分析的基本功能为“一体”。

由于偏最小二乘回归在建模的同时实现了数据结构的简化，因此，可以在二维平面团上对多维数据的特性进行观察，这使得在偏最小二乘回归分析中对各个因素的影响进行分析。在一次偏最小二乘回归分析计算后，不但可以得到多因变量对多自变量的回归模型，而且可以在两维平面因上直接观察两组变量之间的相关关系，以及观察样本点间的相似性结构。这种高维数据多个层面的可视见性，可以使数据系统的分析内容更加丰富，同时又可以对所建立的回归模型给予许多更详细深入的实际解释。此外，偏最小二乘方法适应多因变量对多自变量的回归建摸分析，比对逐个因变量做多元回归更加有效，其结论更加可靠，整体性更强。偏最小二乘回归分析的这些将非模型方式的数据认识性分析方法和优化模型方法集中起来的特点及多因变量建模功能正适合均匀设计试验结果数据分析和优化模型的建立。因此，PLS 回归分析建模技术将为均匀设计的更广泛应用提供有力的技术支持。

参考文献

[1] 方开泰，1980，均匀设计，应用数学学报, 3, 363-372.
[2] 方开泰，1994，均匀设计及其应用，数理统计与管理, 13, 57-63.
[3] 方开泰，1994，均匀设计与均匀设计表，北京：科学出版社
[4] 方开泰、马长兴，2001，正交与均匀试验设计，北京：科学出版社
[5] 方开泰、王元，1996，数论方法在统计中的应用，北京：科学出版社
[6] 唐启义，http//www.statforum.com（网站）
[7] 唐启义、冯明光,2002,实用统计分析及其 DPS 数据处理系统,北京:科学出版社
[8] 王惠文，1999，偏最小二乘回归方法及其应用，北京：国防工业出版社
[9] 张承恩，http://ust40.html.533.net（网站）