DOI: 10.13656/j.cnki.gxkx.2003.02.006
广西科学 Guangxi Sciences 2003, 10 (2): 101~103

共线性在多元二次多项式回归模型中的危害及其处理方法

Problems Caused by Collinearity and Solution in the Multivariate Second-degree Polynomial Models

廖 森 王建设 陈超球*
Liao Sen Wang Jianshe Chen Chaoqiu

(广西大学化学化工学院 南宁市大学路 100号 530004)
(College of Chemistry and Chemical Engineering, Guangxi Univ., 100 Daxuelu, Nanning, Guangxi, 530004, China)

摘要 以 1个均匀设计应用的实例为对象,回归分析得到 2个多元二次多项式方程: (1) $Y = 0.084\,826 + 0.231\,79X_3 - 0.050\,286X_3X_3 + 0.028\,422X_1X_3 - 0.001\,396\,2X_2X_3$; (2) $Y = 0.062\,320 + 0.251\,1X_3 - 0.060\,0X_3X_3 + 0.023\,47X_1X_3$。对这 2个方程进行线性变换处理后,用 SPSS软件包进行共线性诊断分析,得出方程 (1) 第 4项和第 2项存在强烈的共线性关系,共线性的存在使方程的预测变得不可靠; 方程 (2) 没有共线性存在,可以正确指导科研工作。

关键词 多元二次多项式回归模型 共线性 回归分析
中图法分类号 O212

Abstract The two multivariate second-degree polynomial models, (1) $Y = 0.084\,826 + 0.231\,79X_3 - 0.050\,286X_3X_3 + 0.028\,422X_1X_3 - 0.001\,396\,2X_2X_3$, (2) $Y = 0.062\,320 + 0.251\,1X_3 - 0.060\,0X_3X_3 + 0.023\,47X_1X_3$, are obtained from an application of uniform design. These two models transformed by linear method, are analyzed by SPSS Kit to solve collinearity problems. The strong collinearity relation between the fourth item and the second item in Equation (1) is found, which leads to fallibility prognosis. There is no collinearity found in Equation (2) which could be employed in scientific researches.

Key words multivariate second-degree polynomial model, collinearity, regression analyse

在对均匀设计试验的数据进行处理时,多元回归分析是 1种必不可少的数理统计工具[1,2]。近年来,人们对多元线性回归模型中多重共线性的存在、所产生的问题及其处理方法进行了广泛的研究[3~10]。均匀设计对数据进行回归分析时广泛采用的是多元二次多项式模型。多元二次多项式模型是一种非线性的模型,人们往往忽视这种模型中共线性的存在,故这种模型中是否有共线性的存在,共线性将产生什么危害,如何处理共线性问题,均未见有相关的报道。多元二次多项式模型可以进行线性转换,在这种模型中的共线性问题是有可能存在的,也需要给予重视,本文以一具体的例子对该问题进行初步探讨。

2002-10-09收稿, 2002-12-23修回。

* 广西师范学院化学系 南宁市明秀东路 19号 530001 (Dept. of Chemistry, Guangxi Teachers College, 19 Mingxiudonglu, Nanning, Guangxi, 530001, China)。

1 多元二次多项式模型中共线性的存在

1.1 回归分析

表 1[11]的数据经多元逐步回归分析[11],可以得到如下 2个回归方程:

(1) $Y = 0.084\,826 + 0.231\,79X_3 - 0.050\,286X_3X_3 + 0.028\,422X_1X_3 - 0.001\,396\,2X_2X_3$.

相关的参数: $F_1 = 131.493\,7$, $F_2 = 69.200\,4$, $F_3 = 101.345\,0$, $F_4 = 10.364\,3$, $F = 138.224\,6$, $R = 0.998\,2$, $e = 0.010\,71$, $F_0 = 2.1$.

方程及方程的第 1至第 3项均通过 $T = 0.01$的 F检验,方程的第 4项通过 $T = 0.05$的检验,故方程

表 1 制备阿魏酸的均匀设计试验方案 U_7 (7^3) 和结果

Table 1 The U_7 (7^3) and results of uniform design for preparing ferulic acid

No.	配比 Matching (X_1)	吡啶量 Pyridine (X_2)	反应时间 Reaction time (X_3)	收率 Recovery (Y)
1	1.0	13	1.5	0.330
2	1.4	19	3.0	0.360
3	1.8	25	1.0	0.294
4	2.2	10	2.5	0.476
5	2.6	16	0.5	0.209
6	3.0	22	2.0	0.451
7	3.4	28	3.5	0.482

显著。

(2) $Y = 0.062\,320 + 0.251\,1X_3 - 0.060\,0X_3X_3 + 0.023\,47X_1X_3$.

相关的参数: $F_1 = 41.078\,7$, $F_2 = 31.827\,2$, $F_3 = 23.839\,5$, $F = 43.879\,2$, $R = 0.988\,8$, $^e = 0.021\,74$, $F_0 = 2.2$.

方程及方程的各项均通过 $T = 0.01$的 F检验，故方程显著。

由上可见回归分析并没有告诉我们 2个方程中的各项之间是否存在着共线性的关系。要想进一步获得方程中各项间是否存在共线性关系，需要对方程进行共线性分析。

1.2 多元二次多项式回归方程的共线性诊断分析

1.2.1 对方程进行线性变换

多元二次多项式方程进行线性变换处理后也可以用 SAS, SPSS或者 MINTAB等常见的统计分析软件包进行共线性诊断分析。现以上述的方程 (1)为例说明变换的方法。

原方程变换后得方程:

$Y = 0.084\,826 + 0.231\,79Z_1 - 0.050\,286Z_2 + 0.028\,422Z_3 - 0.001\,396\,2Z_4$.

方程经过线性变换后，接着用原始实验的数据去变换成 Z_1, Z_2, Z_2, Z_4所对应的数据，变换的方法见表 2

1.2.2 共线性诊断分析

在 SPSS软件上用表 2中 Z与 Y的数据进行多元线性回归分析，而且选择共线性诊断的功能项进行共线性诊断分析，结果见表 3 用类似的变换方法对方程 (2) 进行共线性诊断，结果列于表 4

表 3 表 4中 Model B Std. Error t VIF分别代表方程的各项(包括常数项)、各项的系数、各项的标准误差、各项的 t统计量、各项(不包括常数项)的 VIF值；Condition index为方程的条件因子。VIF与

表 2 数据变换结果

Table 2 Data tranform results

No.	X_1	X_2	X_3	Z_1 (X_3)	Z_2 (X_3X_3)	Z_3 (X_1X_3)	Z_4 (X_2X_3)	Y
1	1.0	13	1.5	1.5	2.25	1.5	19.5	0.330
2	1.4	19	3.0	3.0	9.0	4.2	57	0.360
3	1.8	25	1.0	1.0	1.0	1.8	25	0.294
4	2.2	10	2.5	2.5	6.25	5.5	25	0.476
5	2.6	16	0.5	0.5	0.25	1.3	8	0.209
6	3.0	22	2.0	2.0	4.0	6.0	44	0.451
7	3.4	28	3.5	3.5	12.25	11.9	98	0.482

表 3 方程 (1) 共线性诊断结果

Table 3 Results of collinearity diagnostics for Equation (1)

模型 Model	B	标准误差 Std. Error	t	VIF	条件因子 Condition index
C	8.472E-02	0.019	4.353		
Z_1	0.231	0.021	10.929	24.951	
Z_2	-5.09E-02	0.06	-8.044	37.377	
Z_3	2.956E-02	0.03	10.000	5.880	
Z_4	-1.44E-03	0.00	-3.182	9.125	32.674

表 4 方程 (2) 共线性诊断结果

Table 4 Results of collinearity diagnostics for Equation (2)

模型 Model	B	标准误差 Std. Error	t	VIF	条件因子 Condition index
C	6.143E-02	0.036	1.695		
Z_1	0.251	0.041	6.817	22.748	
Z_2	-6.10E-02	0.011	-5.529	28.070	
Z_3	2.444E-02	0.005	4.903	5.880	26.606

条件因子均是共线性是否存在的指标值，均不宜大于 30 由表 3可见，方程 (1) 中第 4项 (Z_4) 对标准误差的贡献为 0,即该项对方程的影响不显著，第 2项 (Z_2) 与第 4项 (Z_4) 的 t统计量与其它变量的项的 t统计量相比均较小，故第 4项 (Z_4) 是不必要的。由于第 4项 (Z_4) 进入了方程，引起第 2项 (Z_2) VIF值异常增大，也引起整个方程的条件因子 (32.674) 异常的增大，并超过了 30, 说明第 4项 (Z_4) 与第 2项 (Z_2) 有着强烈的共线性关系，进一步回归分析时须把第 4项去掉。

表 4是方程 (1)去掉第 4项 (Z_4)，消除共线性影响后所得到的结果。由表 4可见，每一项对标准误差的贡献均大于 0,而且各变量项的 VIF值均小于 30, 方程的条件因子 (Condition index= 26.606)小于 30 故第 4项 (Z_4) 消除掉后方程解除了共线性的影响，各变量项间已经没有共线性存在。

以上的分析结果表明在多元二次多项式回归方程中同样存在着共线性的问题

2 多元二次多项式模型中共线性的危害

如果不进行共线性检验，仅从 F 检验值、相关系数 R 以及标准偏差 e 这三项指标考虑，人们往往会选择方程 (1)。而方程 (1) 恰恰是有共线性存在的。由表 5可见，由于共线性的干扰，方程 (1) 的预报值明显偏高，比文献 [1] 值高了近 6个百分点，相对误差则超过 1倍。说明共线性的存在会使方程的预测变得不可靠。而方程 (2) 由于没有共线性的存在，故与文献值相一致。

表 5 寻优结果

Table 5 Results of optimization

方 程 Equation	寻优结果 Optimization results			
	X_1	X_2	X_3	Y
(1)	3.4	10	3.1263	0.5764
(2)	3.4	–	2.575	0.5185
文献 [1] 值 Rom Reference [1]	3.4	–	2.575	0.5185

由表 6可见共线性也导致方程相应项的 F统计值异常的小，即方程 (1)的第 4项的 F_4 仅为 10.3643 通不过 $T=0.01$ 的 F 检验。但与多元线性回归分析有所区别的是，多元二次多项式回归分析中，共线性的存在会使整个方程的显著性、相关系数以及标准偏差等指标变好，呈现出一种过拟合的状态。没有经验的人极易于受到误导，抛弃方程 (2) 转而选择方程 (1)，结果所用的方程往往对实际的科研工作起到误导作用，得不到正确的结果。

表 6 方程 (1) 与方程 (2) 的比较

Table 6 Comparison between Equation (1) and Equation (2)

方 程 Equation	F_1	F_2	F_3	F_4
(1)	131.4937	69.2004	101.3450	10.3643
(2)	41.0787	31.8272	23.8395	–

方 程 Equation	F	R	e	共线性 Collinearity
(1)	138.2246	0.9982	0.01071	有 Yes
(2)	43.8792	0.9888	0.02174	无 No

3 讨论

均匀设计在化学化工的各领域中已经得到广泛深入地应用，可是作者在与一些人士就均匀设计的应用进行交流时，他们常反映说，由均匀设计所得的回归方程常常预报不准，有时甚至与验证试验的结果相去甚远。之所以产生这样的情况，原因在于：① 有些人对回归分析没有足够的理解，在进行非线性回归时片面追求大的 R 值或者小的 e 值，致使选进方程中的项过多，使误差自由度为 1，甚至为 0，此时有关的结论就没有可靠性了，故应使误差有足够的自由度，以自由度≥ 5[11] 为好；② 方程有共线性存在，使方程的可靠性变差甚至没有可靠性。因此，要想在科研实践中正确地应用均匀设计，在进行回归分析时，必须考虑并消除上述两种原因所造成的影响。

本文用回归分析相关的软件包得到回归方程后，再用 MINTAB或者 SPSS软件进行共线性分析，若发现有共线性的存在，则把方程中有共线性关联项中 F 统计值或者 t 统计值最小的那一项人为地去掉，再进行逐步回归分析得到新的方程，接着对新的方程再进行共线性检验，直至找到没有共线性或者共线性程度较小的方程为止。比如，方程 (1) 的第 4项 (Z_4) 不但对标准误差的贡献为 0，而且 t 统计量也是变量项中最小，故须把第 4项 (Z_4) 人为地去掉，接着再进行逐步回归分析得到新的方程，即方程 (2)，对方程 (2) 进行共线性诊断得出方程 (2) 已经没有共线性存在，即可选择方程 (2) 进行回归分析。

本文的方法也适用于其他多元多次多项式回归模型的共线性分析及处理。

参考文献

1 方开泰.均匀设计与均匀设计表.北京：科学出版社，1994.

2 曾昭钧.均匀设计及其应用.沈阳：辽宁人民出版社，1994.

3 范立新.回归分析中多重共线性诊断方法.国外医学 (卫生学分册)，1994, 21(1): 34~37.

4 王惠文，朱韵华. PLS回归在消除多重共线性中的作用.数理统计与管理，1996, 15(6): 48~52.

5 孟庆和.多元回归分析中多重共线性的处理.中国卫生统计，1997, 14(1): 49~50.

6 蔡增正.回归模型中的多项共线性问题.生产力研究，1998, (5): 33~36.

7 林华珍，倪宗瓒.多重共线性变量的回归系数估计及检验.中国公共卫生，1999, 15(2): 131~132.

8 王 伟，田庆伟.建立回归模型应注意避免多重共线性.数理医药学杂志，1999, 12(4): 309~310.

9 刘旭阳.多元共线性引发的问题及其实际的处理方法.湖北大学学报 (自然科学版)，1996, 18(4): 333~337.

10 程龙生，吴可法，黄志同.消除复共线性影响的一种新的解决办法.工程数学学报，1998, 15(2): 113~119.

11 周纪芗.实用回归分析方法.上海：上海科学技术出版社，1990.

(责任编辑：邓大玉)