GMDH与PLS解决多重共线性问题的比较研究

贺昌政[1],吕 欣[2]

(1.四川大学 工商管理学院,成都 610064;2.国防科技大学 信息系统与管理学院,长沙 410073)

摘 要:本文通过理论分析、数据试验以及实证研究三种途径,对GMDH与PLS两种算法解决多重共线性问题的特点进行了比较分析,发现了使用GMDH对于解决多重共线性数据建模问题的贡献,为解决多重共线性问题提供了新的途径。

关键词:多重共线性;GMDH;偏最小二乘法

中图分类号:O212.6 **文献标识码**:A **文章编号**:1002-6487(2007)08-0004-03

0 引言

在普通多元线形回归的应用中,我们常受到许多限制,最典型的问题就是自变量之间的多重共线性。如果采用普通的最小二乘方法,这种变量多重共线性就会严重危害参数估计,扩大模型误差,并破坏模型的稳定性。为了解决多重共线性对模型造成的不良后果,长期以来科研人员进行了大量研究,提出了各种不同的解决办法,如利用逐步回归法删除不重要解释变量、变量一阶差分变换、岭回归分析、主成分分析等。然而这些方法均存在一些不足之处[1]。

伍德(S.Wold)和阿巴诺(C.Albano)等人于1983年首次提出偏最小二乘(Partial Least Squares,PLS)回归。作为一种新型的多元数据分析方法,近十年来,它在理论、方法和应用方面都得到了迅速的发展。密西根大学(Michigan University)的弗耐尔(Fornell)教授称偏最小二乘回归为第二代回归分析方法。

在对多变量系统中的信息进行辨识和筛选的过程中,它的建模策略综合了主成分分析和典型相关分析的思想,在多变量 $x_1,\cdots,x_p$ 中逐次提取综合成分 $t_1,\cdots,t_m(m<p)$,从而将自变量系统中的信息重新组合,有效地提取对系统解释性最强的综合变量,摒除重叠信息或无解释意义的信息干扰,从而克服变量多重共线性在系统建模中的不良作用,得到一个更为准确可靠的分析结果。

数据分组处理方法GMDH[2](Group Method of Data Handling)是近年来发展起来的一种数据挖掘方法,它最早由乌克兰科学院A.G.Ivakhnenko院士于1969年提出[3]。GMDH网络本来是Ivakhnenko(1971)为预报海洋河流中的鱼群提出的模型,又成功地应用于超音速飞机的控制系统(Shrier,1987)和电力系统的负荷预测(Sagara和Murata,1988)。GMDH是建立在人类生存历史中最古老的、最富有成效的试探法则-选择学说基础之上的,它将黑箱思想、生物神经元方法、归纳法和GÖdel的数理逻辑方法有机地结合起来。对于有噪声的小数据样本,它通过建立非物理模型,能给出较准确的拟合与过程预测。

1 GMDH解决多重共线性机理分析

作为自组织数据挖掘的核心算法,GMDH方法采用多项式神经元网络,从根本上避开了多重共线情况导致最小二乘估计结果的诸多问题。由文[4],GMDH方法的停止法则由最优复杂度原理给出:当模型的复杂度逐渐增加时,具有"外补充"性质的称之为外准则的准则值达到极小,全局极小的实现标示最优复杂度模型的存在。GMDH算法是通过不能再改善外准则值停止的,其停止法则可以保证在一定噪声水平下得到数据拟合精度和预测能力之间实现最优平衡的最优复杂度模型[5]。

GMDH建模的特点是数据分组和贯穿于整个建模过程中的内、外准则的运用。它将观测样本数据分为训练集(training set)和检测集(testing set)。由于应用不完全归纳法[6],它能从被研究对象的众多影响因素中筛选出最具有相关性的一些输入变量并生成简洁的模型结构,所以任何一个可能影响被研究对象的变量都可以当作潜在的输入变量而不必考虑多重共线性问题。因此,GMDH算法选取的变量(因素)可尽量全面、广泛,不必经过专门的主观筛选,客观信息容量大。

2 数据试验及结果分析

为比较两种算法在处理多重共线性问题时的差异,下面,全面考虑不同的共线性水平,通过数据试验,分别进行偏最小二乘和GMDH建模。其中偏最小二乘算法的实现使用笔者自行编制的Matlab6.0源程序,GMDH实现使用KnowledgeMiner5.0,数据都经过标准化处理:$x_{ij}^{*}=\frac{x_{ij}-\bar{x}_j}{s_j}$。

(注:定义 $MAPE=\frac{\sum|y-\hat{y}|}{N}$, KnowledgeMiner定义 PESS=

基金项目:国家自然科学基金资助项目(70271073)

$\frac{1}{N}\sum_{t=1}^{N}(y_t-f(x_t,a_t))^2$)[3]

2.1 试验一：完全共线，无噪声数据试验

构造一组数据如下：x_1, x_2, y，其中，$x_2=2x_1$，$y=x_1+x_2$，x_1 由随机函数产生。

本组试验做了多次，试验结果均表现为 PLS 与 GMDH 方法能得到回归方程，而且能 100%地拟合数据。PLS 模型包含了 x_1, x_2，解释能力强；而 GMDH 模型只使用了一个自变量。

2.2 试验二：完全共线，含噪声数据试验

系统真实模型：$y_0=f(x)+\alpha_i Z$，$f(x)=3.6+x_1+2.2x_2-0.8x_3$ 其中，$x_2=2.8x_1$，$x_3=x_1+x_2$。Z 为均匀分布的噪声数据，x_1，Z 由随机函数产生。每组数据 20 个样本，共 2000 条数据。

将数据标准化后，根据 $\alpha_i=0.1、0.5、1$，把此组试验累计做 100 次，统计试验结果见表 1。

表 1　受到噪声影响的完全共线变量数据试验结果比较

α	建模次数	拟合效果(MAPE)				预测效果(PESS)			
		占优的次数		占优的比例		占优的次数		占优的比例	
		PLS	GMDH	PLS	GMDH	PLS	GMDH	PLS	GMDH
0.1	34	34	0	100%	0%	32	2	94.12%	5.88%
0.5	33	34	0	100%	0%	33	0	100%	0%
1.0	33	33	0	100%	0%	33	0	100%	0%
累计	100	100	0	100%	0%	98	2	98%	2%
平均值(α	PLS	2.05%	10.00%	19.75%		PLS	0.07%	1.56%	6.01%
=0.1,0.5,1.0)	GMDH	2.44%	12.08%	23.83%		GMDH	0.07%	1.64%	6.36%

注：此处的 PESS 值由 PLS 之定义转换为 GMDH 之定义即 PESSPLS/N=PRESSGMDH

2.3 试验三：高度共线数据试验

本组高度共线的数据来源于科奈尔（Cornell）1990 年采用的一个化工方面的例子[1]。此后，又被武德（Wold）、德昂赫斯（Tenenhaus）等人多次引用，成为单因变量偏最小二乘回归的经典案例。该例中，有 7 个自变量 x_1~x_7，因变量记为 y。

将数据导入程序，标准化后，应用偏最小二乘算法，提取了两个成分：

$$t_1=-0.4369x_1^*-0.0348x_2^*-0.4372x_3^*-0.3694x_4^*+0.2589x_5^*+0.5129x_6^*-0.3876x_7^*$$

$$t_2=0.1232x_1^*-0.6841x_2^*+0.1278x_3^*-0.1687x_4^*-0.3319x_5^*+0.5698x_6^*+0.2169x_7^*$$

计算得到的最终结果为：

$$y=0.4825t_1+0.269t_2$$

$$=-0.1777x_1^*-0.2008x_2^*-0.1766x_3^*-0.2236x_4^{**}+0.0357x_5^*+0.4007x_6^*-0.1287x_7^*$$

GMDH 输入输出模型：

$$y=0.000020+0.871375x_6^*-0.171846x_7^*$$

模型精度结果比较如表 2 所示。

表 2　回归结果比较

	偏最小二乘法	GMDH
PESS(预测误差平方和)	0.028925	0.0211
SS(拟合误差平方和)	0.2553	0.170854
MAPE(平均绝对误差)	0.1009857	0.0998915

2.4 数据试验结果分析

上述试验说明，使用 GMDH 和 PLS 方法对共线性数据进行建模，不再存在普通最小二乘情况下系数无法估计。模型精度和稳定性不能保证的问题，即两种方法均能解决多重共线性数据建模问题。

在数据完全共线以及存在一定噪声的情况下，使用 PLS 的精度要比 GMDH 高，在数据高度共线的情况下，GMDH 比 PLS 表现出更高的拟合能力和预测精度。

从模型系数的解释能力来看，PLS 最终得到整个自变量空间的线性组合，即能对所有变量进行全面、合理的经济解释，而 GMDH 仅使用极少的自变量个数，突出了重要变量的作用，对自变量空间高度概括。

本文虽然研究了完全共线情况下两者建模差异，但应该注意的是，实际生活尤其是复杂的经济系统中，这种变量之间完全共线的情况是极少出现的，因此，试验一和试验二的研究结论仅仅从理论上说明了 PLS 和 GMDH 具有解决完全共线性的能力。

试验三给建模者以启发，那就是在建模过程中，当发现自变量数据存在高度共线性时，考虑使用 GMDH 方法建立模型以获得更高的拟合精度和预测精度，找出重要变量，使用 PLS 方法建立模型以对整个自变量空间进行合理的经济学意义上的解释。两者的建模结果可以结合起来，相辅相成，从而比较完善地解决多重共线性对建模造成的危害。

基于这种启发，本文使用上述思路，将从实证的角度，研究我国实际利用外资额与经济发展因素的关系。

3　我国实际利用外资额影响因素分析

3.1 数据说明

改革开放以来，越来越多的外资进入中国市场，为研究其与我国经济的联系，借鉴文[7]的指标选取，取 1989-2004 年我国相关经济指标如下：

①AUFC-实际利用外资额；②GDP-国内生产总值；③ECC-国家财政基本建设支出；④TAXI-税率；⑤IMP-进口总额；⑥EXP-出口总额；⑦DEBT-国内外债务总额；⑧TOUI-国际旅游收入；⑨RESC-国家外汇储备；⑩LC-劳动成本；⑪FUND-金融机构人民币存贷款余额；⑫CPI-通货膨胀率。

数据来源：《中国统计年鉴 2005(~2003)》；《中华人民共和国年鉴 2005》；《中国对外经济统计年鉴 2004》；《中国旅游统计年鉴 2005》；《中国商务年鉴 2005》；《关于 1989(~1999)年国家预算执行情况和 1990(~2000)年国家预算草案的报告》。

由于我国统计制度的变化，1994 年后国际旅游收入采用与国际接轨的新办法，之前数据已经经过校正。通货膨胀率采用居民消费价格指数(CPI)替代，用人均 GDP 来表示劳动力成本。

3.2 数据预分析

为消除不同量纲作用，同样对所有原始数据进行标准化处理，观察变量之间的相关性发现除了 CPI 以外，其余变量之间相关系数大部分在 0.8 以上，更多的为 0.9 以上，将标准化后数据导入 SPSS，经过回归计算及多重共线性分析，GDP 和 FUND 的方差膨胀因子高达 82950 和 10467，大大超出阈值 10，采用进入法(Enter)得到回归模型，SPSS 自动排除了这两个自变量。其他变量的方差膨胀因子除了 CPI 外，也全部大大超出基准，说明数据存在严重的多重共线性。

SPSS 得到的模型校正后复测定系数 $R^2=0.922$，$F=20.765$ 通过检验 Sig.=0.007432。然而观察各个自变量的 t 检验值，

除了 DEBT,FUND,LC 以外,其他 8 个均不能通过(小于显著性水平 Sig.)。这是在多重共线性情况下导致的常见后果。

从系数解释能力来看,IMP 的系数 1.012 为负,进口额与实际利用外资额成正比,亦不符合实际。

由上面的分析可以知道,普通多元回归分析在对多重共线性数据建模时面临许多问题,其回归模型不适用。于是考虑使用 PLS 和 GMDH 建模予以讨论。

3.3 偏最小二乘回归模型

将数据导入程序,标准化后,应用偏最小二乘算法,最终提取了两个成分:

t_1=0.3653GDP+0.3004TAXI+0.3346TOUI+0.3192DEBT+0.2704IMP+0.2883EXP+0.2757RESC+0.3145FUND+0.3725LC+0.2718ECC-0.1368CPI

t_2=0.4905GDP-0.1275TAXI+0.2148TOUI+0.0499DEBT-0.2886IMP-0.1766EXP-0.2588RESC-0.0096FUND+0.5774LC-0.3131ECC+0.3073CPI

计算得到的最终结果为:

AUFC=0.239t_1-1.3058t_2

=0.7291GDP-0.095TAXI+0.3611TOUI+0.1416DEBT-0.313IMP-0.1622EXP-0.2727RESC+0.0626FUND+0.8446LC-0.3447ECC+0.3694CPI

3.4 GMDH 输入输出模型

应用 GMDH 算法最终得到的回归方程为:

AUFC=0.000017+0.088650CPI+2.748616GDP-1.930349TAXI

3.5 建模结果分析

表 3

精度	PLS	GMDH
PESS(预测误差平方和)	1.3094	0.7696
SS(拟合误差平方和)	1.450192	0.588906
MAPE(平均绝对误差)	0.241257	0.144139

由表 3 可见,偏最小二乘法和 GMDH 算法均给出了比较令人满意的回归结果。

从模型的拟合能力来看,两者的平均绝对误差分别为 0.2413 和 0.1441,均在可以接受的范围之内。而且 GMDH 出现了预期的拟合精度比 PLS 高的情况。

从模型的预测能力来看,PLS 的预测误差平方和比其拟合误差平方和还要小,表现出极强的模型预测能力。GMDH 预测误差仍然优于 PLS 模型。

从模型的解释能力来看:

(1)偏最小二乘最终提取了两个成分 t_1,t_2,由于共线性自变量较多,其关系错综复杂,通过提取成分,可以编制综合考评标准,这一点类似于主成分分析方法。PLS 最终的回归结果表示,GDP(国内生产总值)和 LC(劳动力成本,此处为人均 GDP)对因变量起主要作用,它们的系数分别为 0.7291 和 0.8446,说明我国实际利用外资额的提高,与我国廉价的劳动力成本和国内经济的稳定发展和人民生活水平的提高是分不开的,说明在我国现有劳动力资源情况下,稳步发展,提高人民的生活水平,为经济发展营造一个安定,谐和的社会环境,对于吸引外资,加快经济建设步伐,具有至关重要的作用。表 4 清晰地标明了各经济指标与我国实际利用外资额的相关性。

表 4 相关性分析

	CGDP	TAXI	TOUI	DEBT	IMP	EXP	RESC	FUND	LC	ECC	CPI
AUF	++	-	+	+	-	-	-	+	++	-	+

(2)GMDH 方法给出的结果仍然是相当精简的--仅仅用 GDP、CPI、TAXI 三个变量就以优于 PLS 的拟合精度和预测精度描述了因变量 AUFC。建模结果表明,GDP 是对 AUFC 起积极作用的重要经济指标,国内生产总值越高,越能吸引外资,提高实际利用外资额。TAXI 是对 AUFC 起消极作用的重要经济指标,税率严重制约 AUFC 的提高,这一点是符合实际情况的,税率是影响企业利益的重要因素,然而,TAXI 是政府财政收入的命脉,如何把握两者之间的均衡,至关重要。CPI 与 AUFC 略呈正相关,其系数不大,没有强行分析的必要。

(3)综合两者的建模结论:AUFC 的增长与 GDP、TAXI、LC 相关密切,我国经济要发展,融入国际社会,实现经济全球化(GLOBAL),外资的作用至关重要。因此,中央和地方政府在制定经济策略时,应当首要考虑促进 GDP 增长,加快人民生活水平提高以及适量平衡税收水平,少搞形象工程,为增加外资利用额创造条件,为经济发展营造安定、和谐的社会环境。

4 结论

(1)通过上述三个数据试验的回归模型结果比较以及使用两种建模方法对我国实际利用外资额相关因素的综合分析,GMDH 和 PLS 算法,不论从数据的拟合精度还是数据预测精度上,均能很好地解决数据多重共线性问题,从而弥补传统普通最小二乘回归的不足。

(2)从本文大量数据建模分析的过程中可以看出,面对多重共线性,偏最小二乘法的优势在于对自变量系统的综合提取利用及对因变量全面合理的解释能力上,而 GMDH 的优势在于用精简的自变量完成对数据拟合和预测的高精度要求,两者正好结合起来,相辅相成。从而,通过 GMDH 和 PLS 建模方法的结合运用,能够对复杂的经济系统进行更有效的描述,为决策者提供更科学的决策依据。

参考文献:

[1]王惠文.偏最小二乘回归方法及其应用[M].北京:国防工业出版社,1999.

[2]Madala H R, Ivakhnenko A G, Boca Raton. Inductive Learning Algorithms for Complex Systems Modeling[M]. London,Tokyo:CRC Press Inc,1994.

[3]Muller, Johann-Adolf; Lemke, Frank, Self-Organising Data Mining. An Intelligent Approach To Extract Knowledge From Data [M]. Berlin,Dresden,1999.

[4]Yurachkovskiy Y P. Improved GMDH Algorithms for Process Prediction[J]. Soviet Automatic Control c/c of Avtomatika, 1977,10(5):61-71.

[5]Ivakhnenko A G,Mueller J A. Problems of an Objective Computer Clustering of a Sample of Boservations [J]. Soviet Journal of Automoation and Information Sciences c/c of Avtomatika, 1991, 24(1).54-62.

[6]Sarychev A P. An Averaged Regularity Criterion for the Group Method of Data Handling in the Problem of Searching for the Best Regression[J]. Soviet Journal of Automation and Information Sciences c/c of Avtomatika,1990,23(5).24-29.

[7]Zu Kweon Kim. The Allocation and Moviation of Japanese and US Foreign Direct Investment in an Economically Integrated Area:The Case of the European Union[J]. Sam Advanced Management Journal, 2004.

(责任编辑/亦 民)