主成分回归的建模策略研究

王惠文 王劼 黄海军

(北京航空航天大学 经济管理学院，北京 100191)

摘 要：分析了国际上通用的主成分回归的工作原理和失效原因.在此基础上，提出一种新的主成分回归建模策略：①提取所有主成分建立模型；②删除模型中 t 检验不显著的成分；③用 t 检验显著的成分建立最终需要的模型.由于任一主成分的回归系数和 t 检验值以及与其余主成分无关.因此，当采用向后删除变量法时，如果有多个成分 t 检验不显著，则可以将它们同时删除，而无须逐个删除.采用仿真案例对所提出的方法的合理性进行验证.这种新的建模策略可以有效地提取对因变量有较强解释作用的成分，实现在自变量多重相关条件下的回归建模，并且允许在模型中包含所有的原始变量.此外，该方法的成分筛选过程简便，累计计算误差小于偏最小二乘回归等迭代算法.

关 键 词：回归分析；主成分分析；成分；筛选

中图分类号：C 931.1

文献标识码：A **文 章 编 号**：1001-5965(2008)06-0661-04

Modeling strategy of principle component regression

Wang Huiwen Wang Jie Huang Haijun

(School of Economics and Management, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Abstract: When the mechanism and the reason of failure of the classical principal components regression were analyzed, a new strategy of PCR modeling was presented as: ①deriving all components and modeling with all these components; ②exclude all components which were not significant in t-test; ③modeling with the components which were significant in t-test. Proved the regression coefficient and the t-test value of any principal component were unrelated to the other principal components. It was insured that, when applying backward-delete variables law, all the variables which were not significant in t-test test could be deleted together at the same time. It was not necessary to delete them gradually. A simulation study was given to prove the validity of the strategy. The research indicates that the suggested strategy can effectively derive components which are explainable to dependent variables. Modeling under the condition of multicollinearity is enabled, and all the independent variables can be included. The process of suggested variables selection method is simple, and the accumulated error is smaller than that of partial least-squares regression.

Key words: regression analysis; principal component analysis; components; selection

在工程技术、经济管理等研究领域，多元线性回归是应用最为广泛的量化分析技术之一.长期以来，自变量集合中的多重相关性问题却始终是该方法在应用中最主要的限制.为了较完备地描述和分析系统，分析人员往往倾向于比较周到地选取有关的指标.而这样构成的多变量系统常常存在严重的多重相关性，导致最小二乘估计方法(OLS，Ordinary Least-Square)，严重地扩大模型误差并破坏模型的稳健性.

为了突破这一约束，一个方法是删除不太重

收稿日期：2007-05-15

基金项目：国家自然科学基金资助项目(70371007，70521001，70531010)；北京市自然科学基金资助项目(9052006)

作者简介：王惠文(1957－)，女，辽宁大连人，教授，wanghw@vip.sina.com.

要的相关性变量.然而,在多重相关性十分严重的情况下,常用的变量筛选方法所得结论的准确性和可靠性都会受到影响.另一个方法是岭回归分析[1-2].岭回归估计量的质量取决于偏倚系数 c 的选取.但偏倚系数 c 的选取尚无标准的决策准则,只能凭经验判断.

主成分回归(PCR, Principal Components Regression)的原理是用主成分分析提取的主成分与因变量回归建模.由于提取成分时没有考虑到与因变量的联系,因此经常出现主要成分对因变量的解释性不强的情况.

针对PCR的缺陷,文献[3]中提出了偏最小二乘回归(PLSR, Partial Least Squares Regression)方法.该方法通过迭代算法在自变量集合中逐步提取成分,使其一方面尽可能多地携带自变量集合中的变异信息,同时又对因变量有很强的解释能力[4-7].然而,近年来的理论和实践研究都表明,PLSR的成分提取和参数估计依然会受到自变量严重多重相关性的影响[8-9].此外,其迭代算法容易产生较大的累计计算误差[10].作为一种较复杂的回归方法,它也不易被各领域应用人员广泛熟悉和使用.

本文系统分析了经典主成分回归的工作原理和失效原因.并提出一种新的建模策略.与PLSR方法相比,方法可以有效地处理多重相关性问题,并且计算简便、累计计算误差小,容易被熟悉经典多元分析的应用人员接受.

1 PCR的原理和失效原因

为方便起见,设所有的自变量都是标准化的.PCR的工作步骤如下:

1) 对 p 个自变量构成的数据表做主成分分析,得到主成分 $\boldsymbol{F}_j \in R^n$, $j=1,2,\cdots,p$.根据选定的累计贡献率,选取前 m 个主成分.后面的主成分 $\boldsymbol{F}_{m+1},\boldsymbol{F}_{m+2},\cdots,\boldsymbol{F}_p$ 由于携带信息量很少,将不再参与下一步的回归分析.

2) 采用OLS方法,做 $\boldsymbol{F}_1,\boldsymbol{F}_2,\cdots,\boldsymbol{F}_m$ 对因变量 $\boldsymbol{y}$ 的多元线性回归,得到

$$\hat{\boldsymbol{y}}=\beta_1\boldsymbol{F}_1+\beta_2\boldsymbol{F}_2+\cdots+\beta_m\boldsymbol{F}_m$$

3) $\boldsymbol{F}_1,\boldsymbol{F}_2,\cdots,\boldsymbol{F}_m$ 是 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p$ 的线性组合,因此得回归模型:

$$\hat{\boldsymbol{y}}=\alpha_1\boldsymbol{x}_1+\alpha_2\boldsymbol{x}_2+\cdots+\alpha_p\boldsymbol{x}_p$$

显然,相互直交的主成分 $\boldsymbol{F}_1,\boldsymbol{F}_2,\cdots,\boldsymbol{F}_m$ 避免了在参数估计时使用最小二乘法的困难[11].然而,PCR一直存在着大量的失效案例.选取的前 m 个主成分对因变量的解释性很差,甚至完全不能满足回归建模的需要.相反的,后面几个被省略掉的主成分 $\boldsymbol{F}_{m+1},\boldsymbol{F}_{m+2},\cdots,\boldsymbol{F}_p$ 却有可能与因变量有较高的相关性.这是由于PCR的成分提取只在自变量集合中进行,没有考虑所提成分是否与因变量存在相关关系,因此可能提取的第一主成分概括了自变量集合中的最多信息,但对y的解释能力却最弱.而在多元回归建模中,提取成分的目的在于选择对因变量y解释能力最强的成分,并利用成分的直交性进行OLS建模[11].因此,只要将全部主成分作为备选变量,采用OLS中的变量筛选方法,做多元回归,就可以克服PCR的失效问题.此外,运用全部主成分做多元线性回归,其变量筛选的过程将更加简单方便.

2 新的PCR的建模策略

定理1 设相互直交的 p 个 n 维变量 $\boldsymbol{F}=(\boldsymbol{F}_1,\boldsymbol{F}_2,\cdots,\boldsymbol{F}_p)$ 是的中心化变量(即每个变量的均值为0).自变量 $\boldsymbol{F}_1,\boldsymbol{F}_2,\cdots,\boldsymbol{F}_P$ 对 n 维因变量 $\boldsymbol{y}$ 的OLS回归模型中,$\boldsymbol{F}_k$ 的回归系数 b_k 和 t 检验值 t_k 由 $\boldsymbol{F}_k$ 和 $\boldsymbol{y}$ 唯一确定($1\leqslant k\leqslant p$).

证明 记 $\boldsymbol{B}=(\boldsymbol{B}_1,\boldsymbol{B}_2,\cdots,\boldsymbol{B}_P)$ 为OLS回归模型中的回归系数,即 $\boldsymbol{B}=(\boldsymbol{F}'\boldsymbol{F})^{-1}\boldsymbol{F}'\boldsymbol{Y}$.

记 c_{kk} 为 $(\boldsymbol{F}'\boldsymbol{F})^{-1}$ 矩阵中主对角线上第 k 行(或第 k 列)的元素;E 为回归模型的估计标准误差.

由于 $\boldsymbol{F}'_i\boldsymbol{F}_j=0$, $\forall\, i\neq j$, $1\leqslant i\leqslant p$ 因此

$$c_{kk}=1/\|\boldsymbol{F}_k\|^2 \tag{1}$$

由定理1还可以得到:

$$b_k=c_{kk}\cdot\boldsymbol{F}'_k\boldsymbol{y} \tag{2}$$

根据 t 检验值的计算公式,得到:

$$t_k=b_k/\sqrt{E\cdot c_{kk}} \tag{3}$$

结合式(1)~式(3)可以看出,b_k 和 t_k 都是由 $\boldsymbol{F}_k$ 和 $\boldsymbol{y}$ 唯一确定($1\leqslant k\leqslant p$).定理1的结论得证.

定理1说明,主成分 $\boldsymbol{F}_k$ 的回归系数 b_k 和 t 检验值 t_k 与其余主成分无关.因此,当采用向后删除变量法时,如果有多个成分 t 检验不显著,可以将它们同时删除.而无须逐个删除.这使得变量筛选变得简单方便.

定理2 对因变量 $\boldsymbol{y}$ 与中心化的直交自变量 $\boldsymbol{F}_1,\boldsymbol{F}_2,\cdots,\boldsymbol{F}_p$ 回归建模(模型称为全模型).记 $\boldsymbol{F}_k$ 的 t 检验值为 t_k($1\leqslant k\leqslant p$);再在自变量 $\boldsymbol{F}_1,\boldsymbol{F}_2,\cdots,\boldsymbol{F}_p$ 中任意挑选出包含 $\boldsymbol{F}_k$ 在内的 r 个自变量($r<p$),与因变量 $\boldsymbol{y}$ 回归建模(称为子模型).子模型中,$\boldsymbol{F}_k$ 的 t 检验值为 t_k',则必有 $t_k'\leqslant t_k$ 的结

论成立.

证明　记 E 为 P 个自变量模型的估计标准误差,E'为 r 个自变量模型的估计标准误差. 由定理 1 得

$$t_k' = b_k / \sqrt{E' \cdot c_{kk}} \tag{4}$$

由 $r<p$ 可知,

$$E \leqslant E' \tag{5}$$

对比式(3)~式(5),可知 $t_k' \leqslant t_k$.

定理 2 的结论得证.

定理 2 指出,在使用主成分进行回归的过程中,如果采用向后删除变量法,那么随着模型中的变量不断减少,所有变量的 t 检验值都是递减的. 因此,在全模型中 t 检验不显著的主成分,在子模型中就更不显著.

根据上面定理 1 和定理 2 的结论,给出如下建模步骤.

1) 对标准化的数据表 $\boldsymbol{X}_{n\times p}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p)$ 做主成分分析,提取全部的主成分 $\boldsymbol{F}_j \in R^n$,$J=1,2,\cdots,p$;

2) 以 $\boldsymbol{F}_j \in R^n$,$j=1,2,\cdots,p$ 为自变量,对因变量 $\boldsymbol{y}$ 做 OLS 回归;

3) 在全模型中一次删除所有 t 检验不显著(不能通过 t 检验)的变量. 记通过 t 检验的 r 个自变量为 $\boldsymbol{F}_{i_1},\boldsymbol{F}_{i_2},\cdots,\boldsymbol{F}_{i_r}$;

4) 以 $\boldsymbol{F}_{i_1},\boldsymbol{F}_{i_2},\cdots,\boldsymbol{F}_{i_r}$ 为自变量,对因变量 $\boldsymbol{y}$ 做 OLS 回归. 如果在 $\boldsymbol{F}_{i_1},\boldsymbol{F}_{i_2},\cdots,\boldsymbol{F}_{i_r}$ 中有的变量不能通过 t 检验,则重复第 3)和 4)步,直至所有模型中的自变量全都通过 t 检验,从而得到最终模型;

5) 记最终模型中的自变量为

$$\boldsymbol{F}^{(1)},\boldsymbol{F}^{(2)},\cdots,\boldsymbol{F}^{(s)} \qquad s \leqslant p$$

即　　$\hat{\boldsymbol{y}}=\beta_1\boldsymbol{F}^{(1)}+\beta_2\boldsymbol{F}^{(2)}+\cdots+\beta_s\boldsymbol{F}^{(s)}$

由于成分 $\boldsymbol{F}^{(1)},\boldsymbol{F}^{(2)},\cdots,\boldsymbol{F}^{(s)}$ 是 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots\boldsymbol{x}_p$ 的线性组合,可以得到回归模型:

$$\hat{\boldsymbol{y}}=\alpha_1\boldsymbol{x}_1+\alpha_2\boldsymbol{x}_2+\cdots+\alpha_p\boldsymbol{x}_p$$

如果在步骤 3)中,遇到全部的 p 个主成分都不能通过 t 检验的情况,则说明主成分回归的方法失效. 需尝试其它的建模方法,例如岭回归,PLSR 等;或者考虑采用其它变量作为自变量.

3　仿真案例分析

为验证本文提出的建模策略,采用采用的一个化工方面的数据进行仿真案例研究[7]. 为了说明论文的研究主题,使用了该数据中的自变量集合,如表 1 所示,再根据仿真分析的目的构造因变量 $\boldsymbol{y}$. 在这个自变量集合中存在着高度相关性. 其中,$\boldsymbol{x}_1$ 与 $\boldsymbol{x}_3$ 的相关系数接近 1.

表 1　化工案例原始自变量

样本	$\boldsymbol{X}_1$	$\boldsymbol{X}_2$	$\boldsymbol{X}_3$	$\boldsymbol{X}_4$	$\boldsymbol{X}_5$	$\boldsymbol{X}_6$
1	0	0.23	0	0	0	0.74
2	0	0.1	0	0	0.12	0.74
3	0	0	0	0.1	0.12	0.74
4	0	0.49	0	0	0.12	0.37
5	0	0	0	0.62	0.12	0.18
6	0	0.62	0	0	0	0.37
7	0.17	0.27	0.1	0.38	0	0
8	0.17	0.19	0.1	0.38	0.02	0.06
9	0.17	0.21	0.1	0.38	0	0.06
10	0.17	0.15	0.1	0.38	0.02	0.1
11	0.21	0.36	0.12	0.25	0	0
12	0	0	0	0.55	0	0.37

从表 2 可以看出,对自变量集合进行主成分分析后,前 3 个主成分 $\boldsymbol{F}_1,\boldsymbol{F}_2,\boldsymbol{F}_3$ 的累计贡献率已经高达 92.04%,几乎概括了绝大部分原始变量中的变异信息. 而 $\boldsymbol{F}_5$ 和 $\boldsymbol{F}_6$ 携带的信息很少,它们的贡献率仅为 0.003% 和 0.001%.

表 2　主成分的贡献率

成分	特征值	贡献率	累计贡献率
1	3.411	56.850	56.850
2	1.540	25.671	82.522
3	0.571	9.518	92.040
4	0.477	7.957	99.997
5	0.000	0.003	99.999
6	0.000	0.001	100.000

现将因变量 $\boldsymbol{y}$ 设计成主成分的线性组合:

$$\boldsymbol{y}=b_1\times\boldsymbol{F}_1+b_2\times\boldsymbol{F}_2+\cdots+b_6\times\boldsymbol{F}_6+\varepsilon \tag{7}$$

其中,b_5,b_6 分别取 0.8,0.9;而 b_1~b_4 的取值依次为 0.008,0.01,0.012,0.014,$\varepsilon \sim N(0,0.1)$.

如表 3 所示,$\boldsymbol{F}_5$ 和 $\boldsymbol{F}_6$ 与因变量的相关性远远高于其他主成分. 从表 4 中可以发现,采用全部 6 个主成分建立回归模型,只有 $\boldsymbol{F}_5$,$\boldsymbol{F}_6$ 能够通过 t 检验,而前 4 个主成分 $\boldsymbol{F}_1,\boldsymbol{F}_2,\boldsymbol{F}_3,\boldsymbol{F}_4$ 都不能通过 t 检验.

表 3　各主成分与因变量的相关系数

成分	$\boldsymbol{F}_1$	$\boldsymbol{F}_2$	$\boldsymbol{F}_3$	$\boldsymbol{F}_4$	$\boldsymbol{F}_5$	$\boldsymbol{F}_6$
与 $\boldsymbol{y}$ 相关系数	0.028	0.016	0.011	−0.004	0.674	0.737

表 4　用所有主成分建立的全模型的检验

成分	回归系数	t 检验值	显著系数
$\boldsymbol{F}_1$	0.028	2.485	0.056
$\boldsymbol{F}_2$	0.016	1.359	0.232
$\boldsymbol{F}_3$	0.011	0.974	0.375
$\boldsymbol{F}_4$	−0.004	−0.350	0.741
$\boldsymbol{F}_5$	0.674	58.991	0.000
$\boldsymbol{F}_6$	0.737	64.469	0.000

删除 $\boldsymbol{F}_1,\boldsymbol{F}_2,\boldsymbol{F}_3,\boldsymbol{F}_4$,仅以 $\boldsymbol{F}_5$, $\boldsymbol{F}_6$ 为自变量对因变量回归建模,得到的最终模型的复测定系数

和调整的复测定系数均为0.998，F 检验值等于2438.1，显然通过 F 检验. 表5给出该模型的回归系数和 t 检验值.

将表4中各主成分的回归系数与公式(7)中的系数进行比较，可以看出，回归模型与因变量的设计相符.

表5 以 F_5，F_6 建立的最终模型

变量	回归系数	t 检验值	显著系数
F_5	0.674	47.140	0.000
F_6	0.737	51.517	0.000

得到 y 关于 x 的模型为：

$$y = 88 + 90.2x_1 + 361.3x_2 + 213.7x_3 + 469.3x_4 + 109.4x_5 + 566.2x_6$$

如图1所示，该模型的拟合效果比较理想.

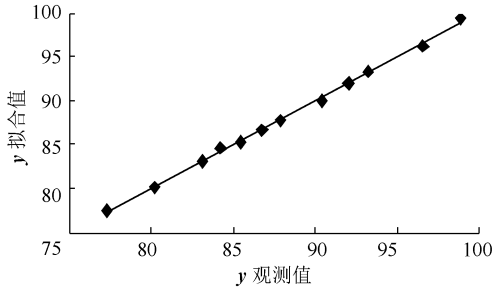


图1 观测值与拟合值的比较图

4 结束语

主成分回归曾经是用于克服多重相关性的一种常用建模方法. 但，它在提取成分时没有考虑与因变量之间的联系. 本文分析了主成分回归的工作原理和失效原因，在此基础上，提出一种新的建模策略，并采用仿真数据对所提出的方法进行验证. 理论研究和仿真分析表明，这种新的主成分建模策略可以有效地实现在自变量严重多重相关条件下的回归建模，而且允许在模型中包含所有的原始变量. 由于该方法采用非迭代算法，因此其累计计算误差显然会小于偏最小二乘回归. 此外，在新的策略下，成分筛选的过程非常简便，计算效率明显高于OLS的变量筛选方法.

参考文献（References）

[1] Hoerl A E, Kennard R W. Ridge regression: biased estimation for non-orthogonal problems [J]. Teehno Metrics, 1970, 12:55－68

[2] Hoerl A E, Kennard R W. Ridge regression:application for non-orthogonal problems[J]. Teehno Metrics, 1970, 12:69－72

[3] Wold S, Albano C, Dunn M, et al. Pattern regression finding and using regularities in multivariate data[M]. London: Analysis Applied Science Publication, 1983

[4] Wold S, Martens H, Wold H. The multivariate calibration problem in chemistry solved by the PLS method[C]// Ruhe A, K gstr m B. Proc Conf Matrix Pencils Lectures Notes in Mathematics. Heidelberg: Springer-Verlag, 1983

[5] Tenenhaus M, L'approche P L S. Revue de Statistique Appliqu e[M]. Paris : Springer-Verlag,1999

[6] Kutner, Nachtsheim, Neter. Applied linear regression models[M]. Fourth Edition. New York: McGraw-Hill,2005

[7] 王惠文. PLSR方法及其应用[M]. 北京：国防工业出版社，1999

Wang Huiwen. Partial least-squares regression method and application [M]. Beijing: National Defence Industry Press, 1999(in Chinese)

[8] Ergon R. Reduced PCR/PLSR models by subspace projections[J]. Chemometrics and Intelligent Laboratory Systems, 2006, 81:68－73

[9] Bjφrn-Helge M, Henrik R C. Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR) [J]. Journol of Chemometrics, 2004, 18: 422－429

[10] Ergon R. Constrained numerical optimization of PCR/PLSR predictors [J]. Chemometrics and Intelligent Laboratory Systems, 2003, 65: 293－303

[11] 任若恩，王惠文. 多元统计数据分析——理论、方法、实例[M]. 北京：国防工业出版社，1997

Ren Ruoen, Wang Huiwen. Statistical analysis on multivariate data-theories, methods, case studies[M]. Beijing: National Defence Industry Press, 1997(in Chinese)

（责任编辑：卢 硕）