偏最小二乘回归法在武器装备研制费用估算中的应用

徐 哲, 刘 荣

(北京航空航天大学经济管理学院, 北京 100083)

摘要: 参数估算法作为LCC的一种估算方法主要应用于武器装备研制费用的估算. 由于样本数据较少, 且变量之间的线性相关程度高, 如果使用传统的普通最小二乘法(OLS)估算, 则会严重降低估计结果的有效性. 而偏最小二乘回归法(PLS)可以克服样本点少和高线性相关性的问题, 使估算结果更趋于真值. 以某型无人侦察机的研制费用估算和美国军用飞机发动机研制费用估算为例, 主要研究偏最小二乘回归法(PLS)在武器装备研制费用估算中的应用.

关键词: 武器装备系统; 参数估算法; 研制费用; 偏最小二乘回归法; 变量投影重要性

1 引 言

60 年代初, 美国国防部提出了全寿命周期费用(Life Cycle Cost, 简称LCC)的概念, 并首先应用在武器采购中. 经过几十年的发展, LCC 已被广泛应用于美国的军事及民用领域. LCC 应包括系统的研制费用、生产费用和保障使用费用[1]. 根据美军的研究, 在典型产品的 LCC 中, 研制阶段实际费用的投入现对较少, 仅占系统LCC 的20% 左右, 但对LCC 的影响却达到了 90% —95%. 因此, 系统研制费用的估算是LCC 的估算和分析中非常关键的一步.

目前国内外用于LCC 估算的方法主要有以下几种[2]: 参数估算法、工程估算法、类比估算法和专家判断法. 其应用范围也不尽相同, 如表 1 所示.

表 1 几种LCC 估算方法的比较

运用阶段 / 估算方法	方案设计	方案论证与确定	早期开发	后期开发	生产制造
参数估算法	*	*	√	×	×
工程估算法	√	√	*	*	*
类比估算法	√	*	×	×	×
专家判断法	√	√	√	√	√

注: * —— 该阶段采用的主要方法; √ —— 该阶段采用的次要方法; × —— 该阶段通常不采用的方法

由表 1 可以看出, 在项目的不同阶段进行费用的估算时, 应该采取相应适宜的方法, 在方案设计与论证阶段通常采用的是参数估算法.

参数估算法也被称为费用估算关系法(Cost Estimating Relationship, CER)[3], 利用同类系统的历史统计数据导出的数学关系来估算新武器系统费用的一种方法, 也是应用较为广泛的一种费用估算方法. 参数法估算是基于系统的物理和性能特点, 并利用旧项目的费

收稿日期: 2004-06-15
基金项目: 航空科学基金资助项目(02J51074)

© 1994-2006 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

用数据得出新项目的值　该方法主要采用回归分析的数学方法, 找出自变量(费用影响因素)与因变量(费用的组成部分)之间的函数关系

运用统计方法进行参数估计时, 样本点越多, 估计出的样本参数越接近总体参数的值　而武器装备系统的样本数据的收集存在着许多困难, 不仅同类产品的历史数据少, 且武器装备又属军事领域, 鉴于保密性, 许多数据均难获得　因此, 当应用普通最小二乘回归法(Ordinary Least-Squares Regression, 简称OLS)进行估计时, 最大的困难就是样本点过少, 使得估计值的准确性很难得到保证

2　偏最小二乘回归法基本原理

通常建立模型时使用的统计方法是OLS, 其优点是计算简单, 易于对变量进行解释　但如果样本点不足, 自变量之间存在严重的多重共线性, 则会使OLS失效, 破坏参数估计, 扩大模型误差, 并使模型丧失稳健性　因此, 我们引入偏最小二乘回归(Partial Least-Squares Regression, 简称PLS)法[4], 与OLS相比, PLS有以下几个特点:

1) 能够在自变量存在严重多重相关性的条件下进行回归建模;

2) 允许在样本点个数少于变量个数的条件下进行回归建模;

3) PLS在最终模型中将包含原有的所有自变量;

4) PLS模型更易于辨识系统信息与噪声(甚至一些非随机性的噪声);

5) 在PLS模型中, 每一个自变量 X_j 的回归系数将更容易解释

2.1　PLS建模原理

设有 q 个因变量 $Y\{y_1, \cdots, y_q\}$ 与 p 个自变量 $X\{x_1, \cdots, x_p\}$, 观测了 n 个样本点　PLS分别在X与Y中提取出成分 t_1 和 u_1(也就是说, t_1 是 $x_1, \cdots, x_p$ 的线性组合, u_1 是 $y_1, \cdots, y_q$ 的线性组合).　在提取这两个成分时, 为了回归分析的需要, 有下列两个要求:

1) t_1 和 u_1 应尽可能大地携带它们各自数据表中的变异信息;

2) t_1 和 u_1 的相关程度能够达到最大

这两个要求表明, t_1 和 u_1 应尽可能好地代表X和Y, 同时自变量的成分 t_1 对因变量的成分 u_1 又有最强的解释能力.　在第一个成分 t_1 和 u_1 被提取后, PLS分别实施X对 t_1 的回归以及Y对 t_1 的回归　如果方程已经达到满意的精度, 则算法终止; 否则, 将利用X被 t_1 解释后的残余信息以及Y被 t_1 解释后的残余信息进行第二轮的成分提取　如此往复, 直到达到一个较满意的精度为止　若最终对X共提取了 m 个成分 $t_1, \cdots, t_m$, PLS将通过施行 y_k 对 $t_1, \cdots, t_m$ 的回归, 然后再表达成 y_k 关于原变量 $x_1, \cdots, x_p$ 的回归方程, $k = 1, 2, \cdots, q$

2.2　PLS建模步骤

记 $E_0 = (E_{01}, \cdots, E_{0p})_{np}$, $F_0 = (F_{01}, \cdots, F_{0q})_{nq}$ 是自变量X与因变量Y经标准化处理后的数据矩阵

1) 记 t_1、u_1 分别是从 E_0、F_0 内提取的第一个成分.

$$t_1 = E_0 w_1, \quad \| w_1 \| = 1$$

$$u_1 = F_0 c_1, \quad \| c_1 \| = 1$$

w_1 是对应于矩阵 $E_0' F_0 F_0' E_0$ 最大特征值的单位特征向量, c_1 是对应于 $F_0' E_0 E_0' F_0$ 矩阵最大特征值的单位特征向量

分别求 E_0 和 F_0 对 t_1, u_1 的三个回归方程

© 1994-2006 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

$$E_0 = t_1 p_1' + E_1, \quad F_0 = u_1 q_1' + F_1^*, \quad F_0 = t_1 r_1' + F_1$$

式中, 回归系数向量是:

$$q_1 = \frac{F_0' u_1}{\| u_1 \|^2}, \quad r_1 = \frac{F_0' t_1}{\| t_1 \|^2}, \quad p_1 = \frac{E_0' t_1}{\| t_1 \|^2}$$

式中, E_1、F_1^*、F_1 分别是三个回归方程的残差矩阵.

2) 用残差矩阵 E_1、F_1 取代 E_0 和 F_0, 进一步求得 w_2 和 c_2, 提取第二个成分 t_2、u_2, 计算回归系数 p_2、r_2, 得到回归方程

$$E_1 = t_2 p_2' + E_2, \quad F_1 = t_2 r_2' + E_2$$

3) 重复以上步骤, 直到回归方程达到满意精度, 这时得到 m 个成分 $t_1, \cdots, t_m$. 则会有

$$E_0 = t_1 p_1' + \cdots + t_m p_m'$$

$$F_0 = t_1 r_1' + \cdots + t_m r_m' + F_m$$

4) 由于 $t_1, \cdots, t_m$ 均可以表示成 E_0 的组合, 因此 F_0 也可以写成 E_0 的线性组合.

5) 按照标准化的逆过程, 将 F_0 的回归方程还原成 Y 对 X 的回归方程.

3 PLS 在样本点过少时的应用

本文以某无人侦察机飞行器分系统研制费用的估算为例[5], 说明 PLS 在费用估算中的应用. 现以研制成功的四种机型的无人侦察机为样本点, 估算飞行器分系统的研制费用.

费用模型为:

$$Y = a_0 X_1^{b_1} X_2^{b_2} X_3^{b_3} X_4^{b_4} X_5^{b_5} X_6^{b_6} X_7^{b_7}$$

对方程两边取对数:

$$\ln Y = \ln a_0 + b_1 \ln X_1 + b_2 \ln X_2 + b_3 \ln X_3 + b_4 \ln X_4 + b_5 \ln X_5 + b_6 \ln X_6 + b_7 \ln X_7$$

取七个飞行器主要技术指标参数作为解释变量. 原始数据(由于保密原因, 已对原始数据进行处理) 见表 2.

表 2 飞行器数据表

样本点	机体研制费用 (万元) Y	续航时间 (小时) X_1	实用升限 (米) X_2	最大平飞速度 (千米/小时) X_3	飞机最大起飞重量 (千克) X_4	无线电控制半径 (千米) X_5	发动机功率 (马力) X_6	技术进步因子 (年) X_7
A	175.435	18	6000	280	560	600	60	13
B	115.528	4	1200	110	420	280	25	3
C	151.063	4	1000	150	720	320	30	12
D	140.553	6	3650	210	400	555	48	5

从原始数据中可以看出解释变量有七个, 而样本点的个数却只有 4 个, 远远少于自变量个数, 用 OLS 无法计算出所需要的结果. 用 PLS 进行回归, 就可以的得出较理想的回归结果. 将表中数据取对数后, 应用统计软件 SMCA—P 进行偏最小二乘回归, 其回归结果为:

$$\ln Y = 2.50908 + 0.0368381 \ln X_1 + 0.0128806 \ln X_2 + 0.0797604 \ln X_3 + 0.19511 \ln X_4 + 0.0469548 \ln X_5 + 0.0598305 \ln X_6 + 0.0837407 \ln X_7$$

而后再进行反对数处理, 所得到费用模型为:

$$Y = 12.29361 X_1^{0.0368381} X_2^{0.0128806} X_3^{0.0797604} X_4^{0.19511} X_5^{0.0469548} X_6^{0.0598305} X_7^{0.0837407}$$

最后对估计值与真实值进行对比分析, 其值如表 3 所示.

© 1994-2006 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 3　估计值与真实值比较分析表

样本点	A	B	C	D
估计值 $\hat{Y}$	176. 23	116. 05	150. 64	139. 67
真实值 Y	175. 44	115. 53	151. 06	140. 55
偏差率 $(\hat{Y} - Y)/Y$	0. 0046	0. 0046	- 0. 003	- 0. 006

从表 3 中, 可以看出, 估计值与真实值之间的偏差比较小, 估计效果比较理想

自变量对因变量的解释还可以用变量投影重要性指标 VIP (Variable Important in Projection) 来测度, 如图 1 所示

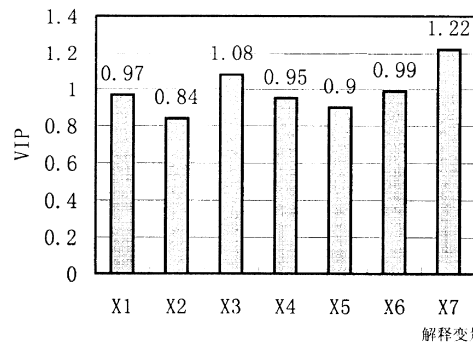

图 1　变量投影重要性柱状图

从图 1 可以看出, 七个解释变量对因变量的解释性都很强, 应全部选取　第七、第三、第六、第一和第四个变量的解释能力最强, 通过回归方程的系数我们也可以发现, x_7 (技术进步因子) 和 x_4 (最大起飞重量) 的值也较其他变量的值大, 这是符合实际情况的　因为飞机的最大起飞重量体现了飞行器分系统的所有结构对费用的影响, 是最重要的性能参数　由于科学技术在这些年的发展进步, 如新材料、新工艺, 特别是电子信息技术的发展日新月异, 对飞行器分系统的研制起了重大影响, 因此模型中技术进步因子也是较大的影响因素　另外, x_1 (续航时间) 和 x_3 (最大平飞速度) 是体现飞行器分系统一个重要性能特征的性能参数, 因而对飞行器分系统的费用影响也比较大

4　PLS 在变量之间存在多重共线性时的应用

PLS 的优点是在建模的过程中通过成分提取的方式, 尽可能多的携带原始数据信息, 弥补了在 OLS 中为了解决多重共线性而牺牲大量信息的缺陷　下面以美国军用飞机从型号设计到型号合格试车的发动机研制费用为例进行了计算分析[6], 其原始数据见表 4 所示

表 4　美国军用飞机发动机技术数据

机型	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
TF30	554. 742	18500	2. 2	2430	3850	51340	0. 6	92	240
TF33	133. 67	17000	1	2060	3900	19240	0. 5	71	458
TF 34	282. 035	9275	1	2660	1420	16500	0. 4	120	338
TF 39	496. 511	40800	1	2840	7300	19500	0. 3	109	1555
J 52	291. 923	8500	1. 8	2060	2050	12840	0. 8	74	122
J 57	199. 897	10000	1. 4	2060	4160	11400	0. 8	41	162
J 60	64. 1915	3000	1	2060	460	10360	1	71	50
J 65	124. 927	7220	1. 2	2030	2815	8500	0. 9	46	117
J 71	252. 066	9570	1. 5	2160	4090	11000	0. 9	47	155
J 75	416. 861	23500	2	2060	5950	16724	0. 8	59	252
J 79	405. 558	15000	2	2160	3225	18056	0. 9	57	162
J 85	330. 418	3850	2	2100	570	10360	1	74	42

© 1994-2006 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

其中, 八个解释变量分别是: X_1 为海平面静止状态最大额定推力 THRMAX (磅); X_2 为最大飞行M 数(与音速有关的速度量)MACH; X_3 为最大涡轮进口温度 TEMP (R); X_4 为发动机净重WGT (磅); X_5 为发动机压力项 TOTPRS (磅/尺^2); X_6 为海平面静止状态最大额定推力下的耗油率 SFCM IL (磅/小时/磅推力); X_7 为到达通过合格试车的时间 TOA (季度); X_8 为最大额定推力下的发动机空气流量 AIR (磅/秒); Y 为从型号设计到型号合格试车的发动机研制费用 COST (百万美元).

变量之间的相关系数如表 5 所示

表 5 变量之间相关系数矩阵

相关系数	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Y
X_1	- 0.054	0.627	0.886	0.381	- 0.666	0.389	0.887	0.648
X_2		- 0.263	0.005	0.409	0.438	- 0.226	- 0.4	0.605
X_3			0.352	0.398	- 0.819	0.846	0.765	0.565
X_4				0.261	- 0.468	- 0.007	0.691	0.502
X_5					- 0.391	0.408	0.116	0.662
X_6						- 0.757	- 0.756	- 0.344
X_7							0.551	0.426
X_8								0.42

从表 5 中可以看出, 自变量之间相关性较大, 存在严重的多重共线性, 如 $r(X_1, X_8) = 0.887$, $r(X_3, X_6) = -0.819$, 如用OLS 对原始数据进行回归, 所得到的回归结果如表 6 所示

表 6 普通最小二乘回归结果

变量	系数	标准误差	t 值	P- 值
常数项	- 838.382	6.733859	- 124.5025	1×10^{-6}
X_1	0.00516	0.000199	25.952825	0.000126
X_2	250.5377	1.319957	189.80743	3×10^{-7}
X_3	0.311526	0.003184	97.84438	2×10^{-6}
X_4	- 0.00059	0.000806	- 0.730036	0.518209
X_5	-7.4×10^{-5}	3.9×10^{-5}	- 1.884759	0.15596
X_6	- 2.89351	4.128233	- 0.700909	0.533834
X_7	- 0.02622	0.054706	- 0.479367	0.664453
X_8	- 0.00094	0.003471	- 0.271898	0.803339

用OLS 回归得到的线性回归方程的通过了 F 检验, 说明因变量与自变量之间存在着显著的线性关系. 但 8 个变量中只有第 1、2、3 个变量的回归系数通过了 t 检验, 而其余各系数均为未通过检验. 这是由于自变量之间存在着多重共线性, 一些变量的信息与其它变量重复了. 而如果为了消除变量之间的多重共线性, 使用逐步回归法进行回归, 其结果只保留了一个变量 X_5, 丢失的信息更多, 回归效果更差. 因此在这里, 我们使用 PLS 来处理, 所有的

© 1994-2006 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

解释变量都对因变量做了程度不同的贡献　通过七次迭代, 所提取的成分覆盖了原始数据 99. 9% 的信息量, 回归结果如下:

$$Y = -822.206 + 0.0039835X_1 + 256.581X_2 + 0.299527X_3 + 0.00179183X_4 - 5.69829 \times 10^{-6}X_5 - 7.83616X_6 + 0.0493524X_7 + 0.02159X_8$$

估计值与真实值的比较见表 7.

表 7　估计值与真实值比较

机型	TF 30	TF 33	TF 34	TF 39	J 52	J 57
真实值 Y	554. 74	133. 67	282. 04	496. 51	291. 92	199. 9
估计值 $\hat{Y}$	555. 21	135. 32	280. 91	496. 97	293. 99	200. 51
残差 $Y - \hat{Y}$	- 0. 466	- 1. 646	1. 1232	- 0. 46	- 2. 062	- 0. 612
机型	J 60	J 65	J 71	J 75	J 79	J 85
真实值 Y	64. 192	124. 93	252. 07	416. 86	405. 56	330. 42
估计值 $\hat{Y}$	61. 177	125. 07	253. 8	414. 24	402. 86	332. 75
残差 $Y - \hat{Y}$	3. 0149	- 0. 147	- 1. 735	2. 6179	2. 7029	- 2. 33

解释变量对因变量的VIP 图见图 2

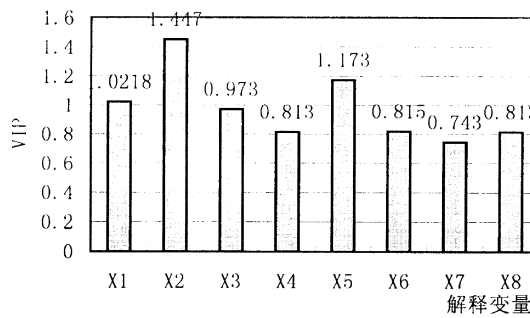

图 2　变量投影重要性柱状图

由图 2 可以看出, 八个解释变量中对因变量贡献最大的变量是飞行速度(即M 数), 其次为压力项、最大推力和涡轮进口温度　对比相关系数矩阵, 容易发现变量的重要程度与自变量同因变量的相关程度基本一致　M 数作为发动机使用环境的指标, 而使用环境是决定所需实验工作量的重要因素, 四分之一以上的研制费与实验有关, 因此M 数是影响研制费用的重要因素　最大推力是发动机重要性能指标之一, 发动机研制费的一半以上用于试验件, 推力的大小反映了试验件的费用, 对研制费的影响是很大的　涡轮进口温度是分析中很重要的变量　其温度的高低反映了该发动机的先进性, 所以如果提高发动机涡轮前的温度就能提高发动机的整体性能, 故此变量是影响研制费用的重要因素之一　压力和耗油率指标在回归方程中的系数为负, 这是因为压力和耗油率在发动机研制过程中要注意控制的两项指标, 则对其要求越高, 指标值就越小, 研制费用也就越高

5　结束语

随着科学技术的发展, 武器的性能不断提高, 对武器装备系统费用的分析也越来越重要　由于研制阶段的费用对于LCC 的重要性, 参数估算法作为该阶段费用的主要估算方法, 其估算结果的准确性对于LCC 的影响很大　将PLS 的统计方法应用于参数估算法来估算研制阶段费用, 克服了OLS 的缺陷, 能够使估算结果更加准确, 对于武器装备研制费用的估算具有重要意义

© 1994-2006 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

参考文献:

[1] 张恒喜. 现代飞机效费分析[M]. 航空工业出版社, 2001.
[2] 顾昌耀. 武器系统全寿命周期费用生成过程的方法研究及应用[J]. 北京航空航天大学经济管理学院, 1990.
[3] 李明, 刘澎等. 武器装备发展系统论证方法与应用[M]. 国防工业出版社, 2000 年 7 月.
[4] 王惠文. 偏最小二乘回归方法及其应用[M]. 国防工业出版社, 1999.
[5] 张嵘. 战场无人侦察机全寿命周期费用估算方法应用研究[C]. 北航硕士论文, 2002.
[6] 飞机费用估算译文集[C]. 航空工业部科学技术情报研究所(译), 1985 年 11 月.
[7] Dr. Gerald R. McNichols, Life cycle cost—art or science[J]. National Aerospace and Electronics conference, MAY-1988.

The Application and Research of PLS in Estimating the Cost of Development in Armament

XU Zhe, LIU Rong

(Beihang University, School of Economy and Management, Beijing 100083, China)

Abstract: Parametric Cost Estimating is a method of analyzing the Life Cycle Cost, which is widely used in estimating the Cost of Development in armament. There are a lot of limitations in the Ordinary Least-Regression if it exits the data scarcity or the high correlation among the variables. Maybe it can cause a bad outcome. If we use the Partial Least-Regression, we can avoid the limitations. The PLS will be introduced at first, and the following are two examples in which the method is put into use.

Keywords: armament system; estimating parameter; cost of development; partial least-regression; variable important in projection

© 1994-2006 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net