Multivariate Data Analysis
Methods And Applications
With Few Observations

小样本多元数据分析方法及应用

张恒喜　郭基联　朱家元　虞健飞　著

西北工业大学出版社

小样本多元数据分析方法及应用

张恒喜　郭基联
朱家元　虞健飞　著

西北工业大学出版社

【内容简介】 本书深入揭示了小样本多元数据的实质和特点，对多元回归法和现代多种建模方法进行了剖析、比较、验证和拓展，提出了小样本多元数据分析的理论和方法，构建了从不同侧面克服小样本多元数据建模困难的完整的建模方法体系。

全书共8章，包括：绪论，多元线性回归分析，偏最小二乘回归分析，方差分量线性模型，自变量筛选和综合特征参数模型，贝叶斯统计分析方法，统计学习理论与支持矢量机，其他分析方法的探讨。

本书可供高等院校飞行器设计、系统工程、管理科学与工程、数量经济学和有关专业的本科生及研究生阅读，也可供研究人员、工程技术人员及有关人员参考。

图书在版编目(CIP)数据

小样本多元数据分析方法及应用/张恒喜等著. —西安：西北工业大学出版社，2002.9

ISBN 7-5612-1561-4

Ⅰ.小… Ⅱ.张… Ⅲ.样本调查(统计学)—多元分析 Ⅳ.0212.2

中国版本图书馆CIP数据核字(2002)第078314号

出版发行：西北工业大学出版社

通信地址：西安市友谊西路127号 **邮编**：710072 **电话**：(029)8493844

网　　址：http://www.nwpup.com

印 刷 者：陕西向阳印务有限公司

开　　本：850 mm×1 168 mm 1/32

印　　张：5.875

字　　数：145千字

版　　次：2002年9月第1版 2002年9月第1次印刷

印　　数：1—2 000册

定　　价：15.00元

前　言

美国学者认为，飞机寿命周期费用参数模型的建模方法迄今为止还没有发现比多元回归更好的方法。这是针对美国的情况而言的。国产飞机寿命周期费用参数模型的建立，有着与美国截然不同的特点。由于国产飞机机型少，费用元素不全和数据时有流失等原因造成样本容量小，多元回归面临困难的状况一直困扰着国产飞机寿命周期费用参数模型的建立。数据和模型的有效性成为人们关注的焦点。能否建立国产飞机寿命周期费用预测模型，不少人持否定态度。有些人则认为至少在现阶段尚不具备条件。有些人对飞机寿命周期费用参数模型的建立表示理解和支持。他们认为"虽然目前建立的模型精度不高，但有总比没有好"。建模方法成为国产飞机寿命周期费用分析的关键技术。

著者自 1987 年起先后承担了"飞机全寿命费用分析"等系列课题的研究工作，一直致力于寻求新的有效的建模方法。经过长时间的探索，考察了国产飞机寿命周期费用历史数据与一般小样本试验数据的区别在于不仅样本容量小，而且具有多维性等特点。在 1998 年承担的国防科技预研基金项目"飞机可靠性/维修性研制与生产费用预测模型研究"中，首先提出了"小样本多元数据"的概念。在对小样本多元数据的实质、特点进行充分分析的基础上，经过对多元回归及众多中外现代建模方法进行剖析、比较、验证和拓展，进一步提出了小样本多元数据分析的理论和方法，成功地构建了从不同侧面克服小样本多元数据建模困难的完整的建模方法体系，应用这些方法建立的国产飞机寿命周期费用预测模型 ALCCM－Ⅲ(2001)，其有效性达到了著名的 DAPCA－Ⅳ(美国

兰德公司用多元回归方法建立）等同类模型的水平。

国产飞机样本容量小的状况将长期存在，因而小样本多元数据分析的理论和方法决不是权宜之计，将长期地得到应用，为数据更新后更新模型发挥积极的作用。本书对小样本多元数据分析理论和方法在管理决策、模式识别等方面的应用也进行了研究。我们相信，小样本多元数据分析理论和方法将不断得到更新和完善。

感谢曾经对我们的工作提供过不同形式支持和帮助的同志们和朋友们。

本书著者有张恒喜（第1、第2、第5章）、郭基联（第3、第4、第5、第8章）、朱家元（第7、第8章）、虞健飞（第6章）。冯惊雷、任博参与了校对工作。

全书由张恒喜统稿。限于著者水平，错误之处敬请读者指正。

著者

2002年9月于西安

目　录

第1章 绪 论

1.1 小样本多元数据分析问题的背景

飞机寿命周期费用(Life Cycle Cost,LCC)的估算是飞机寿命周期费用分析的核心内容和难点。在寿命周期费用估算的不同方法中,参数估算法(Cost Estimate Relation,CER)是应用最为广泛的一种方法。

参数估算法的实质是以一定量的表征工程系统特性的参数(如重量、尺寸、性能、产量、软件程序的行数等)组成费用估算方程或方程组。因此,只要具有在工程系统发展工作中积累起来的类似系统的数据库,就可以建立起费用和工程系统特征量之间的数学关系式。这样,只须输入少量特征参数,即可估算出工程系统在研究、发展、生产阶段的硬件、软件或管理费用,并可达到一定的估算精度。它的另一优点是可以根据估算模型迅速地估计出工程系统的性能或其中某些特征参数的变化对费用的影响(即费用敏感性分析),从而在设计方案的选择及设计方案变更时对费用的影响做出响应。由于参数估算法可用于飞机研制早期阶段,而这一阶段的决策对整个寿命周期费用有重大影响,这就决定了参数估算

法模型在飞机寿命周期费用建模研究中具有十分重要的地位和作用。

美国是最早进行飞机寿命周期费用建模研究的国家，从20世纪60年代初期到80年代中期，建立了包括著名的DAPCA系列参数模型（兰德公司）在内的一批研究成果，并在实践中得到了成功应用。然而，就建模方法而言，这些成果中参数模型的建立方法基本上都采用了多元回归法。这是由于除了受当时建模技术的限制以外，一个根本的原因是多元回归方法的运用需要足够大的样本容量，而美国自莱特兄弟发明第一架飞机以来，航空工业就一直处于全球前列，其优势地位在20世纪70年代以后尤为明显。不论是飞机整体，还是航空发动机、航空电子设备，研制生产所经历的新型号之多、之全在世界各国中是独一无二的；同时，美国又是最早推行寿命周期费用估算的国家，对各型号数据的收集、处理都有一整套相对完备的程序，从而积累了大量的数据。这为多元回归方法的建模研究提供了理论保证。

我国从20世纪80年代中期开始立足于国产飞机进行寿命周期费用估算的研究工作，十几年来已取得了长足的发展。但总的说来，和美国相比，尚未建立一套比较权威的参数模型体系，差距十分明显。这种差距是一种整体上的差距，主要体现在数据的充分性和有效性。事实上，美国由于具备了相对完备的数据库，采用普通的多元回归分析就获得了实用的费用模型，而在我国，数据的收集和整理面临着很大的困难。具体体现在：

（1）国产飞机的型号相对较少且各型号中仿制的型号多，自行研制的少。虽然各机种（或分系统）总的型号数量已相当可观，但具体到某一机种则数量十分有限。如研究战斗机（包括攻击机

和战斗轰炸机）机体研制费用时，美国可用于建模研究的型号有25种，而我国只有10种，且其中有4种是仿制的，其研制费用不能直接用于建模研究；又如美国已经历的脉冲多卜勒制式（PD）机载火控雷达有18种之多，我国仅有4种。这个原因是一个根本性的原因。

（2）费用数据收集工作起步晚，没有建立起完备的数据收集整理程序。数据零散地分布在飞机的研制、生产单位或使用飞机的部队以及机关、院校，尚未建立相对完整的数据库。

（3）对一些厂商来说，出于竞争的需要，产品的研制、生产费用一直被认为是高度机密的商业信息，一般不予公开，虽然在政府对研制工作授予合同或给予资助的情况下，有时可以从官方得到数据。然而经验表明，仅仅参考政府合同的资料不能真实地获得研制、生产的总费用，因为经常是承包商用自己的资金补充政府给予的经费。若仅依赖政府的合同数据，很可能低估真实的费用。这一点在美国也同样存在。

（4）历史上，我国军工产品的研制、生产长期处于计划经济体制下，缺乏按经济规律办事的观念，研制工作常常被当做政治任务，造成有些费用数据不反映实际情况。另外，一些企业内部管理制度不完善，管理水平低下，费用资料比较混乱，残缺不全，发生的费用往往是“一锅煮”，难以分项。

因此，由于数据收集整理上的困难，使得国产飞机寿命周期费用的估算研究成为一个小样本多元数据分析问题，具有和普通多元回归不同的特点和分析方法。

1.2 小样本多元数据分析的特点

1.2.1 小样本多元数据分析的假设条件

从统计的角度而言，多元数据分析一般可分为参数统计和非参数统计。对于一组总体分布形式确定的样本，参数统计总比非参数统计具有更高的功效效率(功效效率的一种简洁的解释是如果使两种模型达到相同的检验功效，样本容量小的则功效效率高)。在国产飞机寿命周期费用建模研究中，由于样本容量小，要想获得较好的估算模型，小样本多元数据分析的研究必须具备一些假设条件。这些假设条件成立与否直接决定了小样本情况下参数模型的可靠性。

(1) 多元回归关系式为线性或可以化为线性关系式(主要表现为对数线性关系式)。

这一点目前是有共识的。一方面，在一定的前提条件下，飞机寿命周期费用模型的形式可以从数理经济学的角度推导出来[1]，主要表现为对数线性关系式；另一方面，从美国现有模型看，他们在反复试算的基础上也选择了对数线性关系式，部分设备直接选择了线性关系式和半对数线性关系式。虽然由于社会制度、经济制度的不同，这些模型不能直接引用，但相似的装备应该具有相似的费用发生机理。

采用线性或对数线性关系式来进行回归分析，并不意味着这一形式就是最合理的。事实上，对于这样的函数逼近一类的问题，线性是相对的，线性问题是非线性问题的特例。这就是在样本容量

充分的条件下，神经网络和模糊系统的非线性函数逼近总会获得较高精度的原因。然而，相比之下，确定一个合理的非线性关系式是比较困难的，这本身需要足够的样本容量。因此，在小样本情况下，根据研究问题的特点事先确定一个确切的关系式形式不仅是合理的，也是必需的。

(2) 费用数据总体服从正态分布，即费用多元线性回归模型的误差服从正态分布。

这是一个比较强的假设。实际数据的总体分布总会或多或少地有所偏离，因为飞机作为一个极为复杂的系统，其费用的发生常会有一些意想不到的因素影响，如政策、国际国内环境等。然而，就整体而言，影响费用发生的因素不仅数目庞大，而且大多是随机和微小的，因此可认为近似服从正态分布。当然，实际研究中须做仔细分析，对于一些受特定因素持久影响(如受政策影响)的样本，其分布已明显偏离了正态分布，成为样本空间中的一个特异点，必须删去或做必要的修正，否则会使模型的特性急剧变差。

误差正态假设是多元线性回归模型假设检验和统计推断的基础。对一组数据，判别其是否服从正态分布对样本容量是有较高要求的，一般应使样本容量 $n \geqslant 50$，这在小样本情况下是不可能的。因此，这一假设条件和前一假设条件一样，对于飞机寿命周期费用的建模研究不仅是合理的，也是必需的。

1.2.2 多元线性回归分析中样本容量问题的讨论

对于一个多元线性回归模型

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e} \tag{1-1}$$

式中，$\boldsymbol{y}$ 是一个 $(n \times 1)$ 的观察值向量；$\boldsymbol{b}$ 是 $[(k+1) \times 1]$ 的回归参

数向量；$\boldsymbol{X}$ 是$[n\times(k+1)]$的自变量常数矩阵；$\boldsymbol{e}$ 是$(n\times 1)$的独立正态随机向量，$\boldsymbol{e}\sim N(0,\sigma_n^2\boldsymbol{I})$；$n$ 为样本容量，k 为自变量个数。

欲得到参数估计值

$$\hat{\boldsymbol{b}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{1-2}$$

则$(\boldsymbol{X}'\boldsymbol{X})^{-1}$必须存在，即$|\boldsymbol{X}'\boldsymbol{X}|\neq 0$。$(\boldsymbol{X}'\boldsymbol{X})$为$(k+1)$阶满秩矩阵，即必须有

$$\mathrm{rank}\boldsymbol{X}\geqslant k+1$$

而 $\boldsymbol{X}$ 为$[n\times(k+1)]$阶矩阵，于是要求

$$n\geqslant k+1 \tag{1-3}$$

也就是说，欲使参数估计值存在，样本容量至少应比模型中的自变量个数多1，这是样本容量的最低限。

作进一步的讨论，当$n\geqslant k+1$时，虽然可以得到参数估计值，但仍然存在这样一些问题：

(1) 参数的统计检验要求样本容量足够地大。

Z 统计量是一种基于标准正态分布的统计量，在总体方差未知时，Z 检验要求 $n\geqslant 30$[5]。t 分布是一种适合于描述小样本特性的分布，即使如此，t 检验作为一种双尾检验，对样本容量也有一定的要求，如图 1-1 所示。由图中可以看出，当显著性水平 $\alpha=0.05$ 时，至少应使 $n-k-1>4$ 时 t 分布变化平缓，检验有效。

(2) 控制假设检验中第 Ⅱ 类错误须满足一定的样本容量。

在假设检验中，当原假设 H_0 为真时做出拒绝 H_0 的错误称为第 Ⅰ 类错误，即“弃真”错误；而当 H_0 为不真时做出接受 H_0 的错误称为第 Ⅱ 类错误，即“取伪”错误。在实际应用中，除了希望控制犯第 Ⅰ 类错误的概率(用α表示)外，往往还希望控制犯第 Ⅱ 类错误的概率(用β表示)。由于犯这两类错误的概率之间存在反比关

系，因而对于任何给定的样本容量，α 的减少将使 β 增大。如果希望同时减少两类错误的可能性，就必须加大样本容量。图 1-2 所示为 $\alpha=0.05$ 时，t 双尾检验的功效（定义为 $1-\beta$）随样本容量变化的曲线。样本取自方差为 σ^2 的正态总体，μ_0 为零假设成立时的平均值。由图可知，犯第 Ⅱ 类错误的概率（β）随样本容量的增加而减少，例如，平均值偏离 μ_0 的值为 σ 时，$N=20$ 则 $\beta\approx0.01$，一个很理想的结果，而 $N=4$ 则 $\beta\approx0.63$，已不能接受。

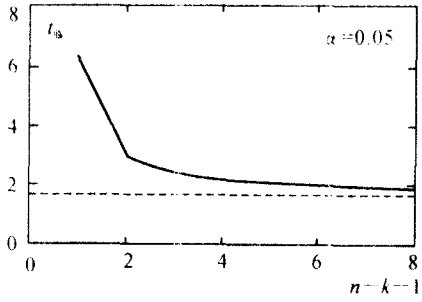

图 1-1　t 临界值随样本自由度的变化

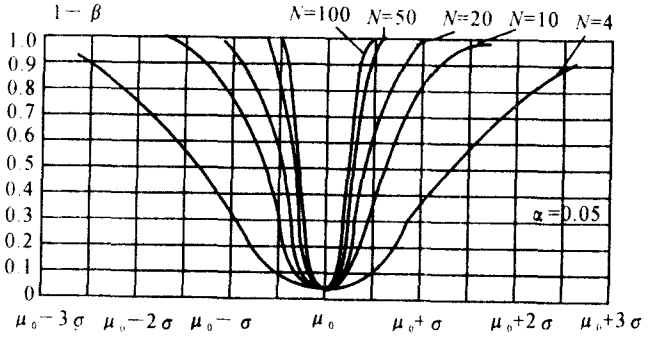

图 1-2　t 双尾检验功效随样本容量的变化

(3) 样本容量不足易使多元线性回归模型产生多重相关性。

一方面，样本容量过小会造成或加剧变量间的多重相关性，此时如果仍采用普通的最小二乘法拟合回归模型，将极大地影响模型的精确性和稳定性。这是因为在式(1-2)中，行列式 $|\boldsymbol{X}'\boldsymbol{X}|$ 趋向于零，$(\boldsymbol{X}'\boldsymbol{X})$ 的逆矩阵会产生很大的舍入误差，严重影响回归系数的值。

另一方面，增加样本容量可以减少线性回归模型参数估计的方差。例如考虑两个自变量的模型

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2}$$

b_1, b_2 的估计方差分别为

$$\begin{aligned} \mathrm{Var}(b_1) &= \frac{\sigma^2}{\sum_{i=1}^{n} x_{i1}^2 (1 - r_{12}^2)} \\ \mathrm{Var}(b_2) &= \frac{\sigma^2}{\sum_{i=1}^{n} x_{i2}^2 (1 - r_{12}^2)} \end{aligned} \tag{1-4}$$

在其他条件不变的情况下，增加样本容量一般会使 $\sum_{i=1}^{n} x_{i1}^2$ 和 $\sum_{i=1}^{n} x_{i2}^2$ 增大，从而使 b_1, b_2 的估计方差减少，提高估计精度。因此，一般认为，在普通多元回归分析中，样本容量至少应是变量个数的两倍以上。

综上所述，对于普通多元线性回归分析，虽然在 $n \geqslant k+1$ 时就可以得到参数估计值，但从参数估计量的有效性来讲，应使 $n \geqslant 30$ 或 $n > k+5$ 或 $n > 2k$(有的文献[8]要求 $n > 3k$)。因此，本书中将不满足这三个条件的一组样本界定为小样本，这样的小样本多元数据的统计问题一般不适于用普通的多元回归分析，而必须另