

\* 学科发展 \*

# 线性模型 M 估计大样本理论的进展与问题

陈希孺\*

(中国科学技术大学研究生院 北京 100039)

**摘要** 文章简要介绍了线性模型 M 估计(包括 LS 和  $L_1$  估计)大样本理论研究状况,提出了一些有待解决的重要问题。

**关键词** 线性模型, M 估计

## 1 引言

线性模型是指形如

$$y_i = x_i' \beta + e_i, \quad i = 1, \dots, n \quad (1)$$

的统计模型,又称线性回归模型,此模型最早起源于 18 世纪天文学家和测地学家的实际工作需要。模型中  $x_i$  ( $p$  维向量)和  $y_i$  是可以直接量测的量,  $\beta$  是根本不可能量测的量,它们根据某种(力学的,天文的,几何的)理论以线性方程  $y = x' \beta$  联系着。在实际观测时有误差,模型(1)中的  $e_i$  就是反映这一项。这模型的基本问题是利用观测数据估计  $\beta$ ,并研究这种估计的精度程度。

1805 年 Legendre 发明最小二乘法(LS 法)。1809 年 Gauss 引进正态误差理论。1823 年 Gauss 证明了 LS 估计的基本性质,即 Gauss-Markov 定理,加上 19 世纪 Helmert、Pizzetti 等人在有关残差平方和的分布理论上的贡献,为这个模型的理论和应用打下了坚实的基础。19 世纪末,统计学家 Pearson 发现此模型与多维正态分布的联系,开辟了把此模型用于社会经济与生物学问题的门径,把当时刚发展起来的相关回归分析纳入此模型的旗下。到 1920 年,统计学家 Fisher 将此模型中自变量取值离散化,使由人工控制试验产生的数据的分析也可以纳入此模型中,其结果就是著名的方差分析法。经过 Gauss、Pearson 和 Fisher 这三位大师的里程碑式的工作,线性模型概括了大部分应用统计的领域,成为数理统计学中最有影响的模型。

此模型在分析统计数据中的实际操作(小样本方法)依赖两个基本点:LS 估计和误差的正态性。二者缺一,则由 Fisher 开辟的小样本方法不能使用。但在实际问题中有时不能坚持这两条。理论研究和应用经验都表明,在一些情况下(如设计矩阵接近病态,数据中含有少量的异常

\* 中国科学院院士  
收稿日期:1998 年 8 月 20 日

值 outlier 等)LS 估计的表现不好。在有些应用(如数量经济)中误差分布属厚尾型,与正态有较大的偏离。为针对 LS 估计可能不适用的情况,研究工作者采取了两个途径:一是仍以 LS 法为基础,加以适当的改进,如 Massy 于 1965 年提出的主成分估计,Hoerl 等于 1970 年提出的岭回归,Webster 等于 1974 年提出的特征根估计等。二是另起炉灶,引进稳定性更好的 M 估计。这个估计把 LS 法中的损失函数由  $u^2$  改为一般的  $\rho(u)$ 。当  $\rho(u) = |u|$  时就是最小一乘 ( $L_1$ )法。它的历史比 LS 法还早,与 LS 估计一起是 M 估计这个家族中最著名的成员。

M 估计,除了 LS 估计,或虽是 LS 估计但误差非正态,都不存在有效的小样本理论,研究工作全集中在大样本(渐近)理论方面。大样本理论是研究当样本量很大时,方法的渐近性态。包括估计量是否收敛于真值(相合性)、渐近分布及其它更深层次的如收敛速度、重对数律、线性表示和渐近展开等问题。现简述有关情况并结合介绍我们自己的一点工作,最后提出一些有待研究的重要问题。

## 2 若干成果

### 2.1 LS 估计的相合性

这方面工作起始于 1960 年代。最初一个重要结果是 Drygas 在 1976 年证明了:在 Gauss-Markov 假设下,LS 估计弱相合的充要条件是

$$S_n^{-1} \rightarrow 0, \quad \text{当 } n \rightarrow \infty, \quad S_n = \sum_{i=1}^n x_i \cdot x_i' \quad (2)$$

$x_i$  的意义见模型(1)。1979 年,美籍统计学家黎子良及 Robbins 和魏庆云证明了:若将 Gauss-Markov 假设加强为误差  $e_1, e_2, \dots$  独立同分布(iid.)并有有限非 0 的方差(此条件还可适当放宽),则(2)是 LS 估计强相合的充要条件。

这些结果在方差有限的条件下基本上解决了误差方差存在时 LS 估计的相合性问题。但从理论上说,“方差存在”对线性模型是一个外加条件。例如在大数律的研究中就不一定作这个要求,而 LS 估计相合性问题实质上是大量律问题的推广,另外从实用角度看,对一些厚尾误差分布也不宜作这一假定,因此有必要研究在方差可能无限时 LS 估计的相合问题。研究集中在:误差  $e_1, e_2, \dots$  iid. 且  $r$  阶矩非 0 有限,  $1 \leq r < 2$ 。在 80 年代有过一些初步结果。到 90 年代,经过我国几位统计学家的努力,问题获得了彻底解决,得到了这种情况下 LS 估计强、弱相合的充要条件。

在研究中还发现了 LS 估计在误差方差无限时的一些特异性质,它从大样本角度显示了:LS 估计的优越性是与误差方差的有限性紧密联系在一起,其中一个反常的性质是:当误差方差不存在时,在模型中添加多余参数反而可能改善 LS 估计的相合性。一般说来,在估计一个参数时,多余参数的存在总会加大估计的困难程度。

此外我们还解决了在 Gauss-Markov 条件下,LS 估计强相合的充要条件问题,此前这问题只有一些初步结果。

### 2.2 $L_1$ 估计的渐近性质

自 50 年代解决了  $L_1$  估计的计算方法问题,计算机性能的提高,以及  $L_1$  估计在某些领域

应用中显示的优越性,60 年代以来这个估计的理论研究得到统计学者的重视。国际上各种刊物发表的结果甚多。主要是在一定的条件下,证明这个估计的相合性与渐近正态性。我国的统计学者也投入了这个领域工作,获得了一些较好的结果,主要是在一些问题上达到了不能改进的条件,或本质上改进了现有的条件。重要的有:

a. 关于  $L_1$  估计的弱相合性,我们在最低的条件(2)之下得到了证明。文献中最好的条件是

$$d_n \equiv \max_{1 \leq i \leq n} x_i' S_n^{-1} x_i \rightarrow 0, \quad \text{当 } n \rightarrow \infty \quad (3)$$

b. 关于  $L_1$  估计的弱相合性,文献中没有一般性的结果。我们证明了:  $d_n = O(1/\log n)$  是充分条件,且对任何常数  $C_n \rightarrow \infty$ , 条件  $d_n = O(C_n/\log n)$  已不再充分。

c. 关于  $L_1$  估计的渐近正态性,国际文献中结果很多,都要求较强的条件  $S_n/n \rightarrow \Lambda > 0$ 。我们将这个条件改进为(3)。这相当于独立和的 Lindeberg 条件,因而已是不可改进的。

### 2.3 一般 M 估计

用于线性模型的 M 估计是 Huber 1973 年提出的。自那以后成为国际统计界研究的一个热点,所得到的成果局限于损失函数  $\rho$  为凸函数的情况。但函数的凸性给其增长速度规定了一个下限,因而对应用是一个较大的限制。我们研究了在损失函数非凸的情况下, M 估计的渐近性质问题,得到了一系列有关线性表示、收敛速度与重对数律等方面比较深刻的结果。另外,在 M 估计的弱相合问题上,在损失为凸时证明了(1)仍是充分条件。对  $x_i$  为随机的情况在  $\rho$  非凸时解决了 M 估计的强相合问题。

## 3 存在的问题

这个领域在 90 年代取得了一系列重要成果,在一定程度上改变了其面貌,但仍有许多重要问题有待解决。首先是损失非  $\rho$  时的问题。目前已有的结果对  $\rho$  施加的条件都太强。如何在对  $\rho$  施加于应用上合理的条件下证明一些基本结果,如相合性与渐近正态性,以及其它更深层次的渐近性质,是在理论上困难且在教育上很有意义的问题。在相合性方面也有不少基本的问题悬而未决,如 M 估计强收敛的条件,弱收敛中条件(1)必要性的证明等,都是数学上的难题。另外,部分线性模型近年来研究的人很多, M 估计是重要题目之一,迄今已得的结果都还很初步。如何把在线性模型研究中积累的一套方法移植于这一模型的研究,以期获得更本质、更具有数学美且在实用上更有意义的结果,是一个富有挑战性的问题。

### 参考文献

- 1 陈希孺,赵林城. 线性模型中的 M 方法. 上海:上海科技出版社,1996.
- 2 Rao C. R., Toutenburg H. Linear Models, Springer Verlag. New York Inc., 1995.