

文章编号: 1000-8055(2004)06-0741-04

不完全数据整体估计方法

傅惠民, 王华胜

(北京航空航天大学 固体力学研究所, 北京 100083)

摘要: 提出一种不完全数据整体估计方法, 给出整体参数的最佳线性无偏估计量及其协方差矩阵, 将传统的只适用于完全数据的回归分析推广到工程中常见的不完全数据的情况。文中针对位置-尺度分布族的不完全数据整体估计方法进行了详细讨论。该方法可以将不同条件下的试验数据作为一个整体进行统计推断, 因此, 对一种条件下只有一个失效数据的情况也能进行分析。与传统方法相比, 具有更高的精度, 而在精度相同的情况下, 则可以节省大量试样。

关键词: 航空、航天推进系统; 不完全数据; 整体估计; 回归分析

中图分类号: TB301; O211

文献标识码: A

Integral Estimate Method for Incomplete Data

FU Huimin, WANG Hua-sheng

(Institute of Solid Mechanics,

Beijing University of Aeronautics and Astronautics, Beijing 100083, China)

Abstract: A new method of integral estimate for incomplete data is presented in this paper. The integral parameter estimators and their covariance are derived. The integral estimate method for the location-scale distributions are discussed at some length. The method extends traditional regression analysis that is only suitable for complete data to incomplete data. Thus it is applicable even when there is only one failure data while the traditional method requires more failure data in a state. As the crosswise information among the incomplete data in different states is exploited and applied, the number of specimens for determining the distribution parameters in all states by the method is about the same as that of the specimens in one state by the traditional method when the precision is equal, which shows the method has the advantage of small sample.

Key words: aerospace propulsion system; incomplete data; integral estimate; regression analysis

回归分析在工程中有着广泛的应用,但是,传统的回归分析只适用于完全数据,对于截尾数据、不完全数据和非正态分布的情况却无法进行处理。然而在工程实际中,如加速寿命试验、生物试

验、外场调查等,由于受试验条件限制和一些无法预料的因素影响,试验中经常出现中止数据,从而形成截尾数据和不完全试验数据。为此,文献[1]建立了一种截尾数据的回归分析,可以将不同条

收稿日期: 2004-09-16; 修订日期: 2004-11-01

基金项目: 国防科技预研项目(413280203)

作者简介: 傅惠民(1956—),男,浙江遂昌人,北京航空航天大学飞行器设计与应用力学系“长江学者”特聘教授,博士,从事小样本信息技术、软校准技术、数据融合方法及可靠性研究

件下的截尾试验数据作为一个整体进行统计推断。文献[2, 3]建立了不完全数据的秩分布和最佳线性无偏估计方法, 解决了单一状态下不完全试验数据的统计分析问题。本文在上述研究工作的基础上, 进一步提出一种不完全数据的整体估计方法, 以提高不完全数据统计推断的精度和节省试样。

1 不完全数据整体估计的基本方程

1.1 多状态下的位置- 尺度分布族

位置- 尺度分布族是工程上常见的一类分布, 如正态分布、Weibull 分布、极值分布、双参数指数分布等, 其分布函数可表示为:

$$F\left(\frac{t - \mu(a, x)}{\sigma(b, x)}\right)$$

概率密度函数为:

$$\frac{1}{\sigma(b, x)} f\left(\frac{t - \mu(a, x)}{\sigma(b, x)}\right), -\infty < t < +\infty$$

$$-\infty < \mu(a, x) < +\infty, \sigma(b, x) > 0$$

其中 $\mu(a, x)$ 和 $\sigma(b, x)$ 为位置参数与尺度参数, $x = (x_1, x_2, \dots, x_h)^T$ 为状态变量, $a = (a_0, a_1, \dots, a_r)^T$ 和 $b = (b_0, b_1, \dots, b_s)^T$ 为整体分布参数。 $\mu(a, x)$ 和 $\sigma(b, x)$ 与 a, b 和 x 之间的函数关系可从所研究问题的物理性质、理论推导、经验公式等多种渠道得到, 也可对其进行检验^[1]。

设 T 为位置- 尺度分布族随机变量, 令 $T^* = [T - \mu(a, x)]/\sigma(b, x)$, 则 T^* 被称为标准位置- 尺度分布族随机变量, 具有分布函数 $F(t^*)$ 和概率密度函数 $f(t^*)$ 。

1.2 位置- 尺度分布族不完全数据顺序统计量

对于来自分布 $F\left(\frac{t - \mu(a, x)}{\sigma(b, x)}\right)$ 的 $n_q + m_q$ 个不完全数据(顺序统计量)的一般形式为:

$$t_1 < t_2 < \dots < t_{n_1} \leq$$

$$t_1^0 \leq t_2^0 \leq \dots \leq t_{m_1}^0 <$$

$$t_{n_1+1} < t_{n_1+2} < \dots < t_{n_2} \leq$$

$$t_{m_1+1}^0 \leq t_{m_1+2}^0 \leq \dots \leq t_{m_2}^0 < \dots <$$

$$t_{n_q-1+1} < t_{n_q-1+2} < \dots < t_{n_q} \leq$$

$$t_{m_q-1+1}^0 \leq t_{m_q-1+2}^0 \leq \dots \leq t_{m_q}^0 \quad (1)$$

式中 $t_k (k = 1, 2, \dots, n_q)$ 表示第 k 个失效数据, $t_k^0 (k = 1, 2, \dots, m_q)$ 表示第 k 个中止数据, $n_j (j = 1, 2, \dots, q)$ 表示前 j 块失效数据总数, $m_j (j = 1, 2, \dots, q)$ 表示前 j 块中止数据总数。令 $t_k^* = [t_k - \mu(a, x)]/\sigma(b, x), k = 1, 2, \dots, n_q$ 和 $t_k^{*0} = [t_k^0 - \mu(a, x)]/\sigma(b, x), k = 1, 2, \dots, m_q$, 则可得到标准位置- 尺度分布 $F(t^*)$ 的 $n_q + m_q$ 个不完全数据(顺序统计量)。

则可得到标准位置- 尺度分布 $F(t^*)$ 的 $n_q + m_q$ 个不完全数据(顺序统计量)。

1.3 不完全数据整体估计的基本方程

设在第 j 个状态 $x_j = (x_{j1}, x_{j2}, \dots, x_{jh})^T$ 下进行寿命试验, 得到 $n_{jq} + m_{jq}$ 个不完全数据, 其中 n_{jq} 个失效数据为 $t_{j1} < t_{j2} < \dots < t_{jn_{jq}}$, 则:

$$t_{jk} = \mu(a, x_j) + \sigma(b, x_j) t_{jk}^* \quad (2)$$

$$j = 1, 2, \dots, m \quad k = 1, 2, \dots, n_{jq}$$

式中 $t_{j1}^* < t_{j2}^* < \dots < t_{jn_{jq}}^*$ 是标准位置- 尺度分布的 n_{jq} 个失效数据。令:

$$\epsilon_{jk} = \sigma(b, x_j) (t_{jk}^* - u_{jk}) \quad (3)$$

式中 $u_{jk} = E(t_{jk}^*)$ 可由文献[3]求得, 将式(3)代入式(2)得到不完全数据整体估计的基本方程:

$$t_{jk} = \mu(a, x_j) + \sigma(b, x_j) u_{jk} + \epsilon_{jk} \quad (4)$$

其均值和协方差分别为:

$$E(t_{jk}) = \mu(a, x_j) + \sigma(b, x_j) u_{jk} \quad (5)$$

$$\text{Cov}\{t_{jk}, t_{jl}\} = v_{jkl} \quad k, l = 1, 2, \dots, n_{jq} \quad (6)$$

式中 $v_{jkl} = \text{Cov}\{t_{jk}^*, t_{jl}^*\}$ 可由文献[3]求得。

2 等尺度分布不完全数据整体估计

工程实际中, 通常可设尺度参数 $\sigma(b, x)$ 为常数, 位置参数 $\mu(a, x)$ 为线性函数, 即:

$$\mu(a, x) = \sum_{k=0}^r a_k \varphi_k(x) \quad (7)$$

$$\sigma(b, x) = \sigma \quad (8)$$

式中 $\varphi_k(x)$ 为已知函数, $k = 0, 1, \dots, r$ 。那么, 由式(4)可得等尺度分布不完全数据整体估计的基本方程:

$$Y = X\theta + \epsilon \quad (9)$$

$$E(Y) = X\theta \quad (10)$$

$$\text{Var}(Y) = V \quad (11)$$

式中 $Y^T = (Y_1^T, Y_2^T, \dots, Y_m^T)$, $Y_j^T = (t_{j1}, t_{j2}, \dots, t_{jn_{jq}})$; $\epsilon^T = (\epsilon_1^T, \epsilon_2^T, \dots, \epsilon_m^T)$, $\epsilon_j^T = (\epsilon_{j1}, \epsilon_{j2}, \dots, \epsilon_{jn_{jq}})$; $\theta = (a_0, a_1, \dots, a_r, \sigma)^T$; $X^T = (X_1^T, X_2^T, \dots, X_m^T)$, 其中:

$$X_j = \begin{bmatrix} \varphi_0(x_j) & \varphi_1(x_j) & \dots & \varphi_r(x_j) & u_{j1} \\ \varphi_0(x_j) & \varphi_1(x_j) & \dots & \varphi_r(x_j) & u_{j2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \varphi_0(x_j) & \varphi_1(x_j) & \dots & \varphi_r(x_j) & u_{jn_{jq}} \end{bmatrix} \quad (12)$$

$$V = \sigma^2 \begin{bmatrix} V_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & V_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & V_m \end{bmatrix} \quad (13)$$

$$V_j = (v_{jkl})_{n_{jq} \times n_{jq}} \quad j = 1, 2, \cdots, m \quad (14)$$

令:

$$Q = (Y - X\theta)^T V^{-1} (Y - X\theta) \quad (15)$$

根据广义高斯-马尔可夫定理可知, θ 的估计量 $\hat{\theta} = (a_0, a_1, \cdots, a_r, \hat{\sigma})^T$ 可由下式给出:

$$\frac{\partial Q}{\partial \theta} = -2X^T V^{-1} (Y - X\theta) = 0 \quad (16)$$

即:

$$\hat{\theta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y \quad (17)$$

可以证明, 式(17)给出 θ 的最佳线性无偏估计, 其协方差为:

$$\text{Cov}(\hat{\theta}) = (X^T V^{-1} X)^{-1} \quad (18)$$

大量Monte Carlo模拟表明, 当数据总量 $\sum_{j=1}^m n_{jq}$ 稍大时, $\hat{\theta}$ 即可近似为正态分布。

可以证明, 在等尺度分布不完全数据情况下, 位置参数 $\mu(a, x)$ 、尺度参数 σ 和百分位值 $\mu(a, x) + F^{-1}(P)\sigma$ 的点估计和置信限估计公式, 与文献[1]给出的截尾数据情况的有关公式相同, 所以

$$X_j = \begin{bmatrix} \varphi(x_j) & \varphi(x_j) & \cdots & \varphi(x_j) & \psi_0(x_j)u_{j1} & \psi_1(x_j)u_{j1} & \cdots & \psi_s(x_j)u_{j1} \\ \varphi(x_j) & \varphi(x_j) & \cdots & \varphi(x_j) & \psi_0(x_j)u_{j2} & \psi_1(x_j)u_{j2} & \cdots & \psi_s(x_j)u_{j2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \varphi(x_j) & \varphi(x_j) & \cdots & \varphi(x_j) & \psi_0(x_j)u_{jn_{jq}} & \psi_1(x_j)u_{jn_{jq}} & \cdots & \psi_s(x_j)u_{jn_{jq}} \end{bmatrix} \quad (23)$$

$$V = \begin{bmatrix} \sigma^2(b, x_1)V_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \sigma^2(b, x_2)V_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \sigma^2(b, x_m)V_m \end{bmatrix} \quad (24)$$

式中 V_j 由式(14)计算。

根据广义高斯-马尔可夫定理可知, θ 的估计量 $\hat{\theta} = (a_0, a_1, \cdots, a_r, b_0, b_1, \cdots, b_s)^T$ 可由下式给出:

$$\hat{\theta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y \quad (25)$$

其协方差为:

$$\text{Cov}(\hat{\theta}) = (X^T V^{-1} X)^{-1} \quad (26)$$

由于式(25)右边含有参数 b , 所以需迭代计算, 初值可取 $b_0 = 1, b_i = 0, i = 1, 2, \cdots, s$ 。

同样, 在非等尺度分布不完全数据情况下, 位置参数 $\mu(a, x)$ 、尺度参数 $\sigma(b, x)$ 和百分位值 $\mu(a, x) + F^{-1}(P)\sigma(b, x)$ 的点估计和置信限估计可采用文献[1]中公式计算。

4 实 例

文献[4]给出了 4 个温度下的某电动机绝缘系统加速寿命试验结果, 每个温度下取 20 个样品, 每

可以采用文献[1]中公式计算。

3 非等尺度分布不完全数据整体估计

工程中有时还会遇到尺度参数随状态变量变化的情况。设位置参数 $\mu(a, x)$ 仍由式(7)给出, 尺度参数 $\sigma(b, x)$ 则为:

$$\sigma(b, x) = \sum_{k=0}^s b_k \psi_k(x) \quad (19)$$

式中 $\psi_k(x)$ ($k = 0, 1, \cdots, s$) 均为已知函数。那么, 由式(4)可得非等尺度分布不完全数据整体估计的基本方程:

$$Y = X\theta + \epsilon \quad (20)$$

$$E(Y) = X\theta \quad (21)$$

$$\text{Var}(\epsilon) = V \quad (22)$$

式中 $\theta = (a_0, a_1, \cdots, a_r, b_0, b_1, \cdots, b_s)^T, X^T = (X_1^T, X_2^T, \cdots, X_m^T)$, 其中:

组试验都出现了中止数据, 得到表 1 中 4 组不完全数据。根据工程经验, 各个温度下电动机绝缘系统的寿命服从 Weibull 分布 $W(\alpha, \beta)$, 且形状参数 α 保持不变, 加速方程满足阿伦尼斯模型, 即:

$$\ln \beta = a + b/T$$

由于 Weibull 分布与极值分布之间存在对数转换关系, 因此, 各个温度下电动机绝缘系统的对数寿命服从极值分布 $EV(\mu, \sigma)$, 尺度参数 σ 保持不变, 且有:

$$\mu = a + b/T$$

采用本文方法对表 1 中不完全数据进行统计分析, 得到:

$$a = -7.527, \quad b = 7214.05, \quad \hat{\sigma} = 0.7647$$

$$\text{Cov} \begin{bmatrix} \hat{a} \\ \hat{b} \\ \hat{\sigma} \end{bmatrix} =$$

$$\begin{bmatrix} 8.258 & -3950.87 & 0.01378 \\ -3950.87 & 1895259.7 & -4.415 \\ 0.01378 & -4.415 & 0.0130 \end{bmatrix}$$

即有:

$$\hat{\beta} = \exp(\hat{a} + \hat{b}/T) = \exp(-7.527 + 7214.05/T)$$

表 1 4 个温度下电动机绝缘系统加速寿命试验结果

Table 1 Accelerated life test results of motor insulation system under four temperature levels

序号	453 K 寿命/h	备注	463 K 寿命/h	备注	493 K 寿命/h	备注	513 K 寿命/h	备注
1	5 606.1	故障	1 628.8	中止	344.1	中止	557.4	中止
2	4 905.1	故障	1 097.7	故障	761.9	中止	156.9	中止
3	2 871.9	中止	630.0	中止	1 562.8	故障	906.5	中止
4	2 763.0	中止	1 520.9	中止	277.0	中止	61.1	故障
5	3 413.8	故障	708.5	中止	482.2	故障	773.4	故障
6	6 321.8	故障	206.0	中止	312.3	故障	148.8	中止
7	4 847.4	故障	185.7	故障	1 434.4	故障	41.2	故障
8	2 690.3	故障	434.3	故障	1 486.6	故障	787.6	故障
9	2 358.2	中止	1 938.7	中止	1 355.5	故障	224.3	中止
10	3 755.4	中止	3 093.8	故障	1 374.0	中止	405.3	中止
11	4 898.8	中止	1 171.9	中止	725.5	故障	1 071.7	中止
12	3 900.2	故障	1 108.8	中止	918.0	中止	407.2	故障
13	1 196.5	故障	27.5	故障	2 970.3	中止	306.0	中止
14	6 000.7	中止	1 428.3	故障	610.0	故障	422.8	故障
15	1 645.2	中止	263.8	中止	89.9	故障	178.6	故障
16	4 021.6	中止	1 113.6	故障	741.6	故障	588.5	故障
17	2 643.7	故障	965.0	中止	706.0	中止	301.6	中止
18	4 760.6	故障	49.1	中止	347.2	故障	14.8	故障
19	1 621.5	故障	350.7	故障	238.6	中止	1 315.0	中止
20	39.0	中止	2 026.7	中止	1 001.4	中止	90.1	故障

$\hat{\alpha} = 1/\hat{\sigma} = 1.3077$

5 结 论

- (1) 提出一种不完全数据整体估计方法, 将只适用于完全数据的回归分析推广到工程中常见的不完全数据的情况。
- (2) 给出整体参数的估计量, 它们是所有条件下不完全数据顺序统计量的线性函数, 具有最佳无偏和渐进正态性质。
- (3) 不完全数据整体估计方法能够将不同条件下的不完全数据作为一个整体进行统计推断, 具有信息量大, 精度高的优点。
- (4) 建立位置- 尺度分布族的不完全数据整体估计的基本方程, 并分别对等尺度分布和非等尺度分布不完全数据整体估计进行了详细讨论。

(5) 非等尺度分布不完全数据整体估计方法可以对工程中常见的异方差数据进行回归分析。

参考文献:

[1] 傅惠民, 黄伟 最佳线性无偏整体估计方法[J]. 机械强度, 2003, 25(3): 319- 324

[2] 傅惠民 不完全数据秩分布理论[J]. 航空学报, 1993, 14(11): 578- 584

[3] 傅惠民, 王华胜 不完全数据最佳线性无偏估计方法[J]. 航空动力学报, 2004, 19(2): 169- 173

FU Huimin, WANG Hua-sheng Best Linear Unbiased Estimate Method for Incomplete Data[J]. Journal of Aerospace Power, 2004, 19(2): 169- 173

[4] Klein J P, Basu A P. Weibull Accelerated Life Tests When There are Competing Causes of Failure[J]. Commun Statist - Theor Meth, 1981, A 10(20): 2073- 2100