

文章编号: 1000-8055(2005)03-0351-04

不完全数据回归分析方法

王华胜, 傅惠民

(北京航空航天大学 固体力学研究所, 北京 100083)

摘要: 基于不完全数据顺序统计量, 提出一种不完全数据回归分析方法。建立不完全数据回归方程, 给出回归系数的最佳无偏整体估计及其协方差矩阵。文中详细讨论了工程中常见的极值分布、Weibull分布和正态分布的等尺度和非等尺度(或异方差)线性回归分析。该方法可以将各个不同状态下的不完全数据作为一个整体进行统计分析, 具有信息量大、精度高的特点。

关键词: 航空、航天推进系统; 不完全数据; 回归分析; 极值分布; Weibull分布; 正态分布

中图分类号: TB301; O211 **文献标识码:** A

Regression Analysis Method for Incomplete Data

WANG Hua-sheng, FU Hu-m-in

(Institute of Solid Mechanics,

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

Abstract: Based on the order statistics of incomplete data, a regression analysis method for incomplete data is presented. The regression equation for incomplete data is established, and the best unbiased integral estimators of the regression parameters and their covariance matrix are also given. The regression analysis methods for incomplete data from extreme distribution, Weibull distribution and normal distribution are discussed in detail. As the cross information among the test data in different states is fully exploited, this present method can reach higher precision than the traditional method.

Key words: aerospace propulsion system; incomplete data; regression analysis; extreme distribution; Weibull distribution; normal distribution

回归分析是一种很重要的数据处理方法, 但是传统回归分析只适用于来自正态分布的完全数据。由于产品失效机理的多样性和寿命影响因素的复杂性, 在工程实际中会遇到各种各样的数据类型和分布规律。为此, 文献[1]建立了截尾数据线性回归分析, 文献[2, 3]分别讨论了不完全数据秩分布理论和不完全数据最佳线性无偏估计方法, 文献[4]提出了不完全数据整体估计方法。本

文进一步给出工程中常见的极值分布、Weibull分布、正态分布不完全数据回归分析方法。

1 极值分布不完全数据回归分析

1.1 极值分布不完全数据顺序统计量

极值分布在航空、航天、军工、船舶、建筑、防灾等工程实际中都有着广泛应用。其分布函数可表示为:

收稿日期: 2004-12-27; 修订日期: 2005-04-10

基金项目: 北京市重点学科工程力学项目(XK100060418)

作者简介: 王华胜(1969-), 男, 山东聊城人, 北京航空航天大学博士生, 从事小样本信息技术、不完全数据分析及可靠性研究

$$F(t) = 1 - \exp\left\{-\exp\left[\frac{t-\mu}{\sigma}\right]\right\}$$

$$-\infty < t < +\infty, -\infty < \mu < +\infty,$$

$$\sigma > 0 \quad (1)$$

式中 μ 和 σ 分别为位置参数和尺度参数。令 $t^* = (t - \mu)/\sigma$, 则 t^* 遵循标准极值分布。

来自极值分布 $EV(\mu, \sigma)$ 的 $n_q + m_q$ 个不完全数据一般可表示为:

$$t_1 < t_2 < \cdots < t_{n_1} \leq t_1^0 \leq t_2^0 \leq \cdots \leq t_{m_1}^0 <$$

$$t_{n_1+1} < t_{n_1+2} < \cdots < t_{n_2} \leq t_{m_1+1}^0 \leq t_{m_1+2}^0 \leq$$

$$\cdots \leq t_{m_2}^0 < \cdots < t_{n_{q-1}+1} < t_{n_{q-1}+2} < \cdots <$$

$$t_{n_q} \leq t_{m_{q-1}+1}^0 \leq t_{m_{q-1}+2}^0 \leq \cdots \leq t_{m_q}^0 \quad (2)$$

其中 $t_k (k = 1, 2, \cdots, n_q)$ 表示第 k 个失效数据或顺序统计量, $t_k^0 (k = 1, 2, \cdots, m_q)$ 表示第 k 个中止数据, $n_j (j = 1, 2, \cdots, q)$ 表示前 j 块失效数据总数, $m_j (j = 1, 2, \cdots, q)$ 表示前 j 块中止数据总数。令 $t_k^* = (t_k - \mu)/\sigma$, 其均值和协方差分别为:

$$E(t_k^*) = u_k = \int \left(\ln \ln \frac{1}{1 - P_k} \right) g(P_k) dP_k$$

$$k = 1, 2, \cdots, n_q \quad (3)$$

$$\text{Cov}(t_k^*, t_l^*) = v_{kl} =$$

$$\iint \left(\ln \ln \frac{1}{1 - P_k} \right) \left(\ln \ln \frac{1}{1 - P_l} \right) g(P_k, P_l) dP_k dP_l$$

$$- u_k u_l \quad k, l = 1, 2, \cdots, n_q \quad (4)$$

式中 $g(P_k)$ 和 $g(P_k, P_l)$ 分别为不完全数据秩统计量的概率密度函数和联合概率密度函数^[3]。

1.2 等尺度极值分布不完全数据回归分析

在工程实际中, 经常遇到极值分布 $EV(\mu, \sigma)$ 的尺度参数 σ 不变, 而位置参数 μ 随状态变量 x 呈线性变化的情况(如各种加速寿命模型), 即:

$$\mu = a + bx \quad (5)$$

此时, 遵循极值分布 $EV(\mu, \sigma)$ 的随机变量 t 可表示为:

$$t = a + bx + \epsilon \quad \epsilon \sim EV(0, \sigma) \quad (6)$$

式中 a , b 和 σ 为待定参数。

设在状态 $x_j (j = 1, 2, \cdots, m)$ 下进行试验, 得到式(2)所示的 $n_{jq} + m_{jq}$ 个不完全数据, 其中有 n_{jq} 个失效数据 $t_{j1} < t_{j2} < \cdots < t_{jn_{jq}}$, 则 a , b 和 σ 的估计量 $\hat{a}$, $\hat{b}$ 和 $\hat{\sigma}$ 由下面诸式给出:

$$\hat{a} = \sum_{i=1}^3 c_i d_{1i} \quad (7)$$

$$\hat{b} = \sum_{i=1}^3 c_i d_{2i} \quad (8)$$

$$\hat{\sigma} = \sum_{i=1}^3 c_i d_{3i} \quad (9)$$

并且 $\hat{\theta} = (\hat{a}, \hat{b}, \hat{\sigma})^T$ 的协方差为:

$$\text{Cov}(\hat{\theta}) = \sigma^2 (d_{kl})_{3 \times 3} \quad (10)$$

式中:

$$c_1 = \sum_{j=1}^m \sum_{k=1}^{n_{jq}} v^{jkl} t_{jl} \quad (11)$$

$$c_2 = \sum_{j=1}^m \sum_{k=1}^{n_{jq}} v^{jkl} x_j t_{jl} \quad (12)$$

$$c_3 = \sum_{j=1}^m \sum_{k=1}^{n_{jq}} u_{jk} v^{jkl} t_{jl} \quad (13)$$

$$(d_{kl})_{3 \times 3} =$$

$$\begin{bmatrix} \sum_{j=1}^m \sum_{k=1}^{n_{jq}} v^{jkl} & \sum_{j=1}^m \sum_{k=1}^{n_{jq}} v^{jkl} x_j & \sum_{j=1}^m \sum_{k=1}^{n_{jq}} u_{jk} v^{jkl} \\ \sum_{j=1}^m \sum_{k=1}^{n_{jq}} v^{jkl} x_j & \sum_{j=1}^m \sum_{k=1}^{n_{jq}} v^{jkl} x_j^2 & \sum_{j=1}^m \sum_{k=1}^{n_{jq}} u_{jk} v^{jkl} x_j \\ \sum_{j=1}^m \sum_{k=1}^{n_{jq}} u_{jk} v^{jkl} & \sum_{j=1}^m \sum_{k=1}^{n_{jq}} u_{jk} v^{jkl} x_j & \sum_{j=1}^m \sum_{k=1}^{n_{jq}} u_{jk} u_{jl} v^{jkl} \end{bmatrix}^{-1} \quad (14)$$

$$(v^{jkl})_{n_{jq} \times n_{jq}} = (v_{jkl})_{n_{jq} \times n_{jq}}^{-1} \quad j = 1, 2, \cdots, m \quad (15)$$

其中 $u_{jk} = E(t_{jk}^*)$, $v_{jkl} = \text{Cov}(t_{jk}^*, t_{jl}^*)$, $k, l = 1, 2, \cdots, n_{jq}$ 为在状态 $x_j (j = 1, 2, \cdots, m)$ 下的标准极值分布不完全数据顺序统计量的均值和协方差, 分别由式(3)和(4)给出。

证明: 根据式(5), 有:

$$t_{jk} = a + bx_j + \sigma t_{jk}^*$$

$$j = 1, 2, \cdots, m, \quad k = 1, 2, \cdots, n_{jq} \quad (16)$$

令 $\epsilon_k = \sigma(t_{jk}^* - u_{jk})$, 则有:

$$E(\epsilon_k) = 0 \quad (17)$$

$$\text{Cov}(\epsilon_k, \epsilon_l) = \begin{cases} \sigma^2 v_{jkl} & i = j \\ 0 & i \neq j \end{cases}$$

$$i, j = 1, 2, \cdots, m$$

$$k = 1, 2, \cdots, n_{jq}, \quad l = 1, 2, \cdots, n_{iq} \quad (18)$$

此时, 式(16)可成为:

$$t_{jk} = a + bx_j + \sigma u_{jk} + \epsilon_k$$

$$j = 1, 2, \cdots, m, \quad k = 1, 2, \cdots, n_{jq} \quad (19)$$

令 $T^T = (T_1^T, T_2^T, \cdots, T_m^T)$, $T_j^T = (t_{j1}, t_{j2}, \cdots, t_{jn_{jq}})$; $\theta = (a, b, \sigma)^T$; $X^T = (X_1^T, X_2^T, \cdots, X_m^T)$; $\epsilon^T = (\epsilon_1^T, \epsilon_2^T, \cdots, \epsilon_m^T)$, $\epsilon_j^T = (\epsilon_{j1}, \epsilon_{j2}, \cdots, \epsilon_{jn_{jq}})$, 则式(19)可写成:

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

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

$$\text{Cov}(T) = \sigma^2 V \quad (22)$$

式中

$$X_j = \begin{bmatrix} 1 & x_j & u_{j1} \\ 1 & x_j & u_{j2} \\ \vdots & \vdots & \vdots \\ 1 & x_j & u_{jn_{jq}} \end{bmatrix} \tag{23}$$

令

$$Q = (T - X\theta)^T V^{-1} (T - X\theta) \tag{24}$$

$$\frac{\partial Q}{\partial \theta} = -2X^T V^{-1} (T - X\theta) = 0 \tag{25}$$

根据高斯-马尔可夫定理可求得 θ 的估计量 $\hat{\theta}$ 为

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

其协方差为:

$$\text{Cov}(\hat{\theta}) = \sigma^2 (X^T V^{-1} X)^{-1} \tag{27}$$

式中

$$V^{-1} = \begin{bmatrix} V_1^{-1} & 0 & \cdots & 0 \\ 0 & V_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & V_m^{-1} \end{bmatrix} \tag{28}$$

$$V_j^{-1} = (v^{jkl})_{n_{jq} \times n_{jq}} \tag{29}$$

由于

$$X^T V^{-1} X = \sum_{j=1}^m X_j^T V_j^{-1} X_j = (d_{kl})_{3 \times 3}^{-1} \tag{30}$$

$$X^T V^{-1} T = \sum_{j=1}^m X_j^T V_j^{-1} T_j = (c_1, c_2, c_3)^T \tag{31}$$

所以, 将式(30)和(31)代入式(26)和(27), 即可得到式(7)~(10), 证毕!

根据文献[4]还可知, $\hat{a}$, $\hat{b}$ 和 $\hat{\sigma}$ 是 a , b 和 σ 的最佳线性无偏估计量, 并且服从渐近正态分布。大量

Monte Carlo 模拟表明, 当数据总量 $n_q = \sum_{j=1}^m n_{jq}$ 稍大时即可用正态分布近似。

1.3 非等尺度极值分布不完全数据回归分析

在工程实际中还常常遇到极值分布 $EV(\mu, \sigma)$

的位置参数 μ 和尺度参数 σ 都随状态变量 x 呈线性变化的情况, 即:

$$\mu = a + bx \tag{32}$$

$$\sigma = c + dx \tag{33}$$

此时, 遵循极值分布 $EV(\mu, \sigma)$ 的随机变量 t 可表示为:

$$t = a + bx + \epsilon \quad \epsilon \sim EV(0, c + dx) \tag{34}$$

式中 a, b, c 和 d 为待定参数。

同理, 对于状态 $x_j (j = 1, 2, \cdots, m)$ 下的 $n_{jq} + m_{jq}$ 个不完全数据, a, b, c 和 d 的估计量 $\hat{a}, \hat{b}, \hat{c}$ 和 $\hat{d}$ 可由下面诸式给出:

$$\hat{a} = \sum_{i=1}^4 c_i d_{1i} \tag{35}$$

$$\hat{b} = \sum_{i=1}^4 c_i d_{2i} \tag{36}$$

$$\hat{c} = \sum_{i=1}^4 c_i d_{3i} \tag{37}$$

$$\hat{d} = \sum_{i=1}^4 c_i d_{4i} \tag{38}$$

并且 $\hat{\theta} = (\hat{a}, \hat{b}, \hat{c}, \hat{d})^T$ 的协方差为:

$$\text{Cov}(\hat{\theta}) = (d_{kl})_{4 \times 4} \tag{39}$$

式中

$$c_1 = \sum_{j=1}^m \sum_{k,l=1}^{n_{jq}} \frac{v^{jkl} t_{jl}}{\sigma_j^2} \tag{40}$$

$$c_2 = \sum_{j=1}^m \sum_{k,l=1}^{n_{jq}} \frac{v^{jkl} x_j t_{jl}}{\sigma_j^2} \tag{41}$$

$$c_3 = \sum_{j=1}^m \sum_{k,l=1}^{n_{jq}} \frac{u_{jk} v^{jkl} t_{jl}}{\sigma_j^2} \tag{42}$$

$$c_4 = \sum_{j=1}^m \sum_{k,l=1}^{n_{jq}} \frac{u_{jk} v^{jkl} x_j t_{jl}}{\sigma_j^2} \tag{43}$$

$$(d_{kl})_{4 \times 4} = \begin{bmatrix} \sum_{j=1}^m \sum_{k,l=1}^{n_{jq}} \frac{v^{jkl}}{\sigma_j^2} & \sum_{j=1}^m \sum_{k,l=1}^{n_{jq}} \frac{v^{jkl} x_j}{\sigma_j^2} & \sum_{j=1}^m \sum_{k,l=1}^{n_{jq}} \frac{u_{jk} v^{jkl}}{\sigma_j^2} & \sum_{j=1}^m \sum_{k,l=1}^{n_{jq}} \frac{u_{jk} v^{jkl} x_j}{\sigma_j^2} \\ \sum_{j=1}^m \sum_{k,l=1}^{n_{jq}} \frac{v^{jkl} x_j}{\sigma_j^2} & \sum_{j=1}^m \sum_{k,l=1}^{n_{jq}} \frac{v^{jkl} x_j^2}{\sigma_j^2} & \sum_{j=1}^m \sum_{k,l=1}^{n_{jq}} \frac{u_{jk} v^{jkl} x_j}{\sigma_j^2} & \sum_{j=1}^m \sum_{k,l=1}^{n_{jq}} \frac{u_{jk} v^{jkl} x_j^2}{\sigma_j^2} \\ \sum_{j=1}^m \sum_{k,l=1}^{n_{jq}} \frac{u_{jk} v^{jkl}}{\sigma_j^2} & \sum_{j=1}^m \sum_{k,l=1}^{n_{jq}} \frac{u_{jk} v^{jkl} x_j}{\sigma_j^2} & \sum_{j=1}^m \sum_{k,l=1}^{n_{jq}} \frac{u_{jk} u_{jl} v^{jkl}}{\sigma_j^2} & \sum_{j=1}^m \sum_{k,l=1}^{n_{jq}} \frac{u_{jk} u_{jl} v^{jkl} x_j}{\sigma_j^2} \\ \sum_{j=1}^m \sum_{k,l=1}^{n_{jq}} \frac{u_{jk} v^{jkl} x_j}{\sigma_j^2} & \sum_{j=1}^m \sum_{k,l=1}^{n_{jq}} \frac{u_{jk} v^{jkl} x_j^2}{\sigma_j^2} & \sum_{j=1}^m \sum_{k,l=1}^{n_{jq}} \frac{u_{jk} u_{jl} v^{jkl} x_j}{\sigma_j^2} & \sum_{j=1}^m \sum_{k,l=1}^{n_{jq}} \frac{u_{jk} u_{jl} v^{jkl} x_j^2}{\sigma_j^2} \end{bmatrix}^{-1} \tag{44}$$

$$(v^{jkl})_{n_{jq} \times n_{jq}} = (v_{jkl})_{n_{jq} \times n_{jq}}^{-1} \tag{45}$$

$j = 1, 2, \cdots, m$

式中 $u_{jk} = E(t_{jk}^*)$, $v_{jkl} = \text{Cov}(t_{jk}^*, t_{jl}^*)$, $k, l = 1, 2, \cdots$, n_{jq} 为在状态 $x_j (j = 1, 2, \cdots, m)$ 下的标准极值分布不完全数据顺序统计量的均值和协方差, 分别由

式(3)和(4)计算。

可以证明,当 σ_j 已知时, $\hat{a}, \hat{b}, \hat{c}$ 和 $\hat{d}$ 是 a, b, c 和 d 的最佳无偏整体估计;当 σ_j 未知时,可先取 $\sigma_j = 1$,由上述公式求得 $\hat{\theta}$ 并根据 $\hat{\theta}$ 计算 σ_j 的估计值 $\hat{\sigma}_j = \hat{c} + \hat{d}x_j$,然后反复迭代计算,直到 σ_j 估计值稳定为止,此时,即可得到满足精度要求的 $\hat{\theta}$ 值。大量 Monte Carlo 模拟表明,当数据总量 $n_q = \sum_{j=1}^m n_{jq}$ 稍大时, $\hat{a}, \hat{b}, \hat{c}$ 和 $\hat{d}$ 近似服从正态分布。

2 Weibull 分布不完全数据回归分析

设 Weibull 分布函数为:

$$F(t) = 1 - \exp\left\{-\left(\frac{t}{\beta}\right)^\alpha\right\} \quad (46)$$

式中 $\alpha > 0$ 为形状参数, $\beta > 0$ 为尺度参数。若令 $y = \ln t, \mu = \ln \beta, \sigma = 1/\alpha$,则式(46)变为极值分布:

$$F(y) = 1 - \exp\left\{-\exp\left[\frac{y - \mu}{\sigma}\right]\right\} \quad (47)$$

式中 μ 和 σ 分别为极值分布的位置参数和尺度参数。由于 Weibull 分布与极值分布存在对数转换关系。因此,对于 Weibull 分布形状参数 α 不变,对数尺度参数 $\ln \beta$ 呈线性变化,即 $\ln \beta = a + bx$ 的情况,可转化为等尺度参数极值分布不完全数据回归分析问题进行处理。

3 正态分布不完全数据回归分析

对于正态分布:

$$f(t) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{(t - \mu)^2}{2\sigma^2}\right\} \quad (48)$$

$$-\infty < t < +\infty, -\infty < \mu < +\infty, \sigma > 0$$

式中 μ 和 σ^2 分别为母体均值和方差。令 $t^* = (t - \mu)/\sigma$,则 t^* 为标准正态随机变量。

设式(2)中的 $n_q + m_q$ 个不完全数据来自正态分布 $N(\mu, \sigma)$,其中有 n_q 个失效数据 $t_1 < t_2 < \dots < t_{n_q}$ 相应的标准正态顺序统计量为 $t_1^* < t_2^* < \dots < t_{n_q}^*$,其均值和协方差分别为:

$$E(t_k^*) = u_k = \int \Phi^{-1}(P_k) g(P_k) dP_k$$

$$k = 1, 2, \dots, n_q \quad (49)$$

$$\text{Cov}(t_k^*, t_l^*) = v_{kl} =$$

$$\iint \Phi^{-1}(P_k) \Phi^{-1}(P_l) g(P_k, P_l) dP_k dP_l - u_k u_l$$

$$k, l = 1, 2, \dots, n_q \quad (50)$$

式中 $\Phi(\cdot)$ 为标准正态分布函数; $g(P_k)$ 和 $g(P_k, P_l)$ 可由文献[3]查得。

用式(49)和(50)代替式(3)和(4),则1.2和1.3节中各个公式可用于正态分布不完全数据的回归分析。

由于寿命数据往往服从对数正态分布,所以可先将不完全寿命数据取对数,然后再按正态分布处理。

4 结 论

(1) 基于不完全数据顺序统计量,提出一种不完全数据回归分析方法,该方法能够全面开发利用各个状态下不完全数据的横向信息,可以提高参数的估计精度。

(2) 建立了等尺度下极值分布不完全数据线性回归方程,给出了回归系数的最佳无偏整体估计及其协方差矩阵。它们同样适用于 Weibull 分布和正态分布的不完全数据线性回归分析。

(3) 给出非等尺度下极值分布不完全数据回归分析方法,有关公式也适用于正态分布不完全数据的线性回归分析。

参考文献

- [1] 黄伟,傅惠民 截尾数据线性回归分析方法[J]. 航空动力学报, 2003, 18(4): 465~469
WUANG Wei, FU Huimin Linear Regression Analysis Method for Censored Data [J]. Journal of Aerospace Power, 2003, 18(4): 465~469
- [2] 傅惠民 不完全数据秩分布理论[J]. 航空学报, 1993, 14(11): 578~584
FU Huimin Theory of Incomplete Data Rank Distribution [J]. ACTA AERONAUTICA ET ASTRONAUTICA SINICA, 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] 傅惠民,王华胜 不完全数据整体估计方法[J]. 航空动力学报, 2004, 19(6): 741~744
FU Huimin, WANG Hua-sheng Integral Estimate Method for Incomplete Data [J]. Journal of Aerospace Power, 2004, 19(6): 741~744