

文章编号: 1000-8055(2003)02-0583-04

威布尔分布多元回归分析方法

黄 伟, 傅惠民

(北京航空航天大学 固体力学研究所, 北京 100083)

摘要: 提出威布尔分布多元回归分析方法, 建立回归参数的最佳无偏整体估计及其协方差公式, 给出威布尔分布、极值分布和正态分布的百分位值(如可靠寿命和安全强度)的置信限估计。传统的多元回归分析只适用于正态分布和完全数据的情况, 而本文则将其推广到威布尔分布、极值分布和截尾数据的情况。与传统的成组试验和最佳线性无偏估计方法相比, 本文方法可以将不同条件的试验数据作为一个整体进行统计推断, 能够全面开发利用不同条件下试验数据之间的横向信息, 在试样数相同的情况下, 具有更高的估计精度, 而在精度相同的条件下, 则可以节省大量试样。

关 键 词: 航空、航天推进系统; 多元回归分析; 最佳无偏整体估计; 截尾试验; 威布尔分布; 极值分布; 正态分布; 可靠性

中图分类号: TB114; O211 **文献标识码:** A

Multiple Regression Analysis Method for Weibull Distribution

HUANG Wei, FU Huimin

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

Abstract: A new multiple regression analysis method for Weibull distribution and extreme distribution is presented. The best unbiased integral estimation for regression parameters and their covariance are established by the method. The point estimators and confidence intervals of Weibull distribution, extreme distribution and normal distribution are also given. The present method extends the traditional multiple regression analysis that is only suitable to normal distribution and complete data to Weibull distribution, extreme distribution and censored data. Compared with the traditional group test method and best linear unbiased estimation, the present method can fully exploit and apply the crosswise information among the test data in different states; thus it reaches a better precision and a great number of specimens can be saved.

Key words: aerospace propulsion system; multiple regression analysis; best unbiased integral estimate; censored test; Weibull distribution; extreme distribution; normal distribution; reliability

收稿日期: 2003-07-09; **修订日期:** 2003-09-22

基金项目: 国防科技预研项目(413280203)

作者简介: 黄伟(1963—), 男, 广西南宁人, 广西大学机械工程学院副教授, 北京航空航天大学固体力学研究所博士生, 从事小样本信息技术、整体推断及可靠性研究

1 引言

通常机电设备需在多种服役环境下工作,工程上要求对每种服役环境下的机电设备进行寿命预测和可靠性评估。由于机电设备的寿命一般服从威布尔分布,而传统回归分析又只适用于来自正态分布的完全数据^[1],所以无法将机电设备在多种服役环境下的试验数据作为一个整体进行统计分析。目前只能采用成组试验法和最佳线性无偏估计方法进行数据处理,但是这两种方法都是对不同条件下的试验数据分别进行处理,这就要求在每种条件下都做大量的试验,从而导致总的试验量很大。为此,本文建立威布尔分布的回归分析方法,将传统的多元回归分析推广到威布尔分布、极值分布和截尾数据的情况,可以对不同条件下的试验数据进行整体推断,既能提高预测精度,又可节省大量试样。

2 威布尔分布多元线性回归分析

2.1 回归系数及其协方差估计

设随机变量 t 服从威布尔分布,则其对数

$y = \ln t$ 服从极值分布,设随机变量 y 与自变量 $x_1, x_2, \dots, x_r$ 成线性关系:

$$l_{i_1 i_2} = \begin{cases} \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} (x_{ji_1} - \bar{x}_{i_1}) (x_{ji_2} - \bar{x}_{i_2}) & 1 \leq i_1 \leq r, \quad 1 \leq i_2 \leq r \\ \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} (x_{ji_1} - \bar{x}_{i_1}) (u_{jk} - \bar{u}) & 1 \leq i_1 \leq r, \quad i_2 = r+1 \\ \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} (u_{jk} - \bar{u}) (u_{jl} - \bar{u}) & i_1 = r+1, \quad i_2 = r+1 \end{cases} \quad (10)$$

$$l_{iy} = \begin{cases} \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} (x_{ji} - \bar{x}_i) (y_{jk} - \bar{y}) & 1 \leq i \leq r \\ \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} (u_{jk} - \bar{u}_j) (y_{jl} - \bar{y}) & i = r+1 \end{cases} \quad (11)$$

其中 $(g_{jkl})_{q_j \times q_j} = (v_{jkl})_{q_j \times q_j}^{-1}$, u_{jkl} 和 v_{jkl} 分别为标准极值分布顺序统计量的数学期望和协方差,由下面诸式计算:

$$u_{jk} = \int \frac{n_j!}{(k-1)!(n_j-k)!} p_k^{k-1} (1-p_k)^{n_j-k} \ln \ln \frac{1}{1-p_k} dp_k \quad (12)$$

$$v_{jkk} = \int \frac{n_j!}{(k-1)!(n_j-k)!} p_k^{k-1} (1-p_k)^{n_j-k} \left(\ln \ln \frac{1}{1-p_k} \right)^2 dp_k - u_{jk}^2 \quad (13)$$

$$v_{jkl} = \iint \left\{ \frac{n_j!}{(k-1)!(l-k-1)!(n_j-l)!} p_k^{k-1} (1-p_l)^{n_j-l} (p_l - p_k)^{l-k-1} \times \right. \\ \left. \ln \ln \frac{1}{1-p_k} \ln \ln \frac{1}{1-p_l} \right\} dp_k dp_l - u_{jk} u_{jl} \quad (k < l) \quad (14)$$

可以证明, $\hat{a}, \hat{b}_1, \hat{b}_2, \dots, \hat{b}_r, \hat{\sigma}$ 是 $a, b_1, b_2, \dots, b_r, \sigma$ 的最佳线性无偏整体估计。大量 Monte Carlo 模拟表

$$y = a + b_1 x_1 + b_2 x_2 + \dots + b_r x_r + \epsilon \\ \epsilon \sim EV(0, \sigma) \quad (1)$$

式中 $a, b_1, b_2, \dots, b_r, \sigma$ 为待定参数。设在状态

$x_j = (x_{j1}, x_{j2}, \dots, x_{jr})$ 进行 n_j 个试样的试验,

$y_{j1} \leq y_{j2} \leq \dots \leq y_{j q_j}$ 为观测到的前 q_j 个定数截尾

样本数据, $j = 1, 2, \dots, m$, 则根据文献[2]可知,

$a, b_1, b_2, \dots, b_r, \sigma$ 的最佳无偏整体估计量 $\hat{a}, \hat{b}_1, \hat{b}_2, \dots, \hat{b}_r, \hat{\sigma}$ 由下面诸式给出:

$$\hat{a} = \bar{y} - \hat{b}_1 \bar{x}_1 - \hat{b}_2 \bar{x}_2 - \dots - \hat{b}_r \bar{x}_r - \hat{\sigma} \bar{u} \quad (2)$$

$$\hat{b} = (\hat{b}_1, \hat{b}_2, \dots, \hat{b}_r, \hat{\sigma})^T = L_{xx}^{-1} L_{xy} \quad (3)$$

式中:

$$n' = \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} \quad (4)$$

$$\bar{x}_i = \frac{1}{n'} \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} x_{ji} \quad (5)$$

$$\bar{u} = \frac{1}{n'} \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} u_{jk} \quad (6)$$

$$\bar{y} = \frac{1}{n'} \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} y_{jk} \quad (7)$$

$$L_{xx} = (l_{i_1 i_2})_{(r+1) \times (r+1)} \quad (8)$$

$$L_{xy} = (l_{iy})_{(r+1) \times 1} \quad (9)$$

明, 当数据总量 $q = \sum_{j=1}^m q_j$ 稍大时, 它们的分布即可近似为正态分布。 $\hat{a}, \hat{b}_1, \hat{b}_2, \dots, \hat{b}_r, \hat{\sigma}$ 的方差、协方差矩阵为:

$$\text{Cov}(\hat{a}, \hat{b}_1, \hat{b}_2, \dots, \hat{b}_r, \hat{\sigma})^T = \sigma^2 (c_{i_1 i_2})_{(r+2) \times (r+2)} \quad (15)$$

式中:

$$(c_{i_1 i_2})_{(r+2) \times (r+2)} = \begin{bmatrix} \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} & \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} X_{j1} & \cdots & \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} X_{jr} & \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} U_{jk} \\ \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} X_{j1} & \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} X_{j1}^2 & \cdots & \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} X_{jr} X_{j1} & \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} U_{jk} X_{j1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} X_{jr} & \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} X_{j1} X_{jr} & \cdots & \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} X_{jr}^2 & \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} U_{jk} X_{jr} \\ \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} U_{jk} & \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} U_{jk} X_{j1} & \cdots & \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} U_{jk} X_{jr} & \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} U_{jk} U_{jl} \end{bmatrix}^{-1} \quad (16)$$

2.2 位置参数和尺度参数的置信限估计

位置参数 $\mu = a + b_1 X_1 + b_2 X_2 + \dots + b_r X_r$ 和尺度参数 σ 的最佳无偏整体估计为:

$$\hat{\mu} = \hat{a} + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \dots + \hat{b}_r X_r \quad (17)$$

$$\hat{\sigma} = \hat{\sigma} \quad (18)$$

μ 和 σ 的置信度为 γ 的单侧置信下限和单侧置信上限分别由下面诸式给出:

$$\mu_L = \hat{a} + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \dots + \hat{b}_r X_r - u_\gamma \hat{\sigma} \sqrt{C_{11} + 2 \sum_{i=1}^r C_{1(i+1)} X_i + \sum_{k,l=1}^r C_{(k+1)(l+1)} X_k X_l} \quad (19)$$

$$\mu_U = \hat{a} + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \dots + \hat{b}_r X_r + u_\gamma \hat{\sigma} \sqrt{C_{11} + 2 \sum_{i=1}^r C_{1(i+1)} X_i + \sum_{k,l=1}^r C_{(k+1)(l+1)} X_k X_l} \quad (20)$$

$$\underline{\sigma} = (1 - u_\gamma \sqrt{C_{(r+2)(r+2)}}) \hat{\sigma} \quad (21)$$

$$\overline{\sigma} = (1 + u_\gamma \sqrt{C_{(r+2)(r+2)}}) \hat{\sigma} \quad (22)$$

2.3 百分位值的置信限估计

随机变量 y 的概率为 P 的百分位值 y_P 的无偏估计量 $\hat{y}_P$ 由下式给出:

$$\hat{y}_P = \hat{a} + \sum_{i=1}^r \hat{b}_i X_i + \hat{\sigma} \ln \ln \frac{1}{1-P} \quad (23)$$

百分位值 y_P 的置信度为 γ 的单侧置信下限 y_{PL} 和单侧置信上限 y_{PU} 分别由下面两式给出:

$$y_{PL} = \hat{a} + \sum_{i=1}^r \hat{b}_i X_i + \hat{\sigma} \ln \ln \frac{1}{1-P} - u_\gamma \hat{\sigma} \left\{ C_{11} + 2C_{1(r+2)} \ln \ln \frac{1}{1-P} + C_{(r+2)(r+2)} \left(\ln \ln \frac{1}{1-P} \right)^2 + 2 \sum_{i=1}^r C_{1(i+1)} X_i + 2 \ln \ln \frac{1}{1-P} \sum_{i=1}^r C_{(i+1)(r+2)} X_i + \sum_{k,l=1}^r C_{(k+1)(l+1)} X_k X_l \right\}^{1/2} \quad (24)$$

$$y_{PU} = \hat{a} + \sum_{i=1}^r \hat{b}_i X_i + \hat{\sigma} \ln \ln \frac{1}{1-P} + u_\gamma \hat{\sigma} \left\{ C_{11} + 2C_{1(r+2)} \ln \ln \frac{1}{1-P} + C_{(r+2)(r+2)} \left(\ln \ln \frac{1}{1-P} \right)^2 + 2 \sum_{i=1}^r C_{1(i+1)} X_i + 2 \ln \ln \frac{1}{1-P} \sum_{i=1}^r C_{(i+1)(r+2)} X_i + \sum_{k,l=1}^r C_{(k+1)(l+1)} X_k X_l \right\}^{1/2} \quad (25)$$

由于任一连续函数均可用多项式来近似, 因此在实际问题中经常碰到回归方程为多项式的情况, 即

$$y = a + b_1 X + b_2 X^2 + \dots + b_r X^r + \epsilon \quad \epsilon \sim EV(0, \sigma) \quad (26)$$

此时, 令 $x_i = x^i, i = 1, 2, \dots, r$, 则式(26)将转化为式(1)的情况, 同样可用上述方法进行回归分析。

3 正态分布截尾数据的多元线性回归

对于随机变量 y 遵循正态分布的情况, 设

$$y = a + b_1 x_1 + b_2 x_2 + \cdots + b_r x_r + \epsilon \quad \epsilon \sim N(0, \sigma^2) \quad (27)$$

此时, 只要将式(12)、(13)和(14)改成下面三式:

$$u_{jk} = \int \frac{n_j!}{(k-1)!(n_j-k)!} p_k^{k-1} (1-p_k)^{n_j-k} \Phi^{-1}(p_k) dp_k \quad (28)$$

$$v_{jk} = \int \frac{n_j!}{(k-1)!(n_j-k)!} p_k^{k-1} (1-p_k)^{n_j-k} [\Phi^{-1}(p_k)]^2 dp_k - u_{jk}^2 \quad (29)$$

$$v_{jkl} = \iint \left\{ \frac{n_j!}{(k-1)!(l-k-1)!(n_j-l)!} p_k^{k-1} (p_l - p_k)^{l-k-1} (1-p_l)^{n_j-l} \Phi^{-1}(p_k) \times \right. \\ \left. \Phi^{-1}(p_l) \right\} dp_k dp_l - u_{jk} u_{jl} \quad (k < l) \quad (30)$$

则式(3)~(11)仍然成立。此时回归曲线及其上、下限分别为:

$$\hat{y} = \hat{a} + \hat{b}_1 \hat{x}_1 + \hat{b}_2 \hat{x}_2 + \cdots + \hat{b}_r \hat{x}_r \quad (31)$$

$$y_L = \hat{a} + \hat{b}_1 \hat{x}_1 + \hat{b}_2 \hat{x}_2 + \cdots + \hat{b}_r \hat{x}_r - u_Y \hat{\sigma} \sqrt{c_{11} + 2 \sum_{i=1}^r c_{1(i+1)} \hat{x}_i + \sum_{k,i=1}^r c_{(k+1)(i+1)} \hat{x}_k \hat{x}_i} \quad (32)$$

$$y_U = \hat{a} + \hat{b}_1 \hat{x}_1 + \hat{b}_2 \hat{x}_2 + \cdots + \hat{b}_r \hat{x}_r + u_Y \hat{\sigma} \sqrt{c_{11} + 2 \sum_{i=1}^r c_{1(i+1)} \hat{x}_i + \sum_{k,i=1}^r c_{(k+1)(i+1)} \hat{x}_k \hat{x}_i} \quad (33)$$

标准差 σ 的单侧置信限仍由式(21)和(22)给出

随机变量 y 的概率为 P 的百分位值 $y_P = a + \sum_{i=1}^r b_i x_i + u_P \sigma$ 的无偏估计量 $\hat{y}_P$ 由下式给出:

$$\hat{y}_P = \hat{a} + \sum_{i=1}^r \hat{b}_i \hat{x}_i + u_P \hat{\sigma} \quad (34)$$

百分位值 y_P 的置信度为 γ 的单侧置信下限 y_{PL} 和单侧置信上限 y_{PU} 分别由下面两式给出:

$$y_{PL} = \hat{a} + \sum_{i=1}^r \hat{b}_i \hat{x}_i + u_P \hat{\sigma} - u_Y \hat{\sigma} \left\{ c_{11} + 2c_{1(r+2)} u_P + c_{(r+2)(r+2)} u_P^2 + 2 \sum_{i=1}^r c_{1(i+1)} \hat{x}_i + \right. \\ \left. 2u_P \sum_{i=1}^r c_{(i+1)(r+2)} \hat{x}_i + \sum_{k,i=1}^r c_{(k+1)(i+1)} \hat{x}_k \hat{x}_i \right\}^{1/2} \quad (35)$$

$$y_{PU} = \hat{a} + \sum_{i=1}^r \hat{b}_i \hat{x}_i + u_P \hat{\sigma} + u_Y \hat{\sigma} \left\{ c_{11} + 2c_{1(r+2)} u_P + c_{(r+2)(r+2)} u_P^2 + 2 \sum_{i=1}^r c_{1(i+1)} \hat{x}_i + \right. \\ \left. 2u_P \sum_{i=1}^r c_{(i+1)(r+2)} \hat{x}_i + \sum_{k,i=1}^r c_{(k+1)(i+1)} \hat{x}_k \hat{x}_i \right\}^{1/2} \quad (36)$$

4 结 论

(1) 提出威布尔分布和极值分布的多元回归分析方法, 给出了回归参数的最佳无偏整体估计公式及其协方差公式。

(2) 给出威布尔分布、极值分布百分位值的置信限估计, 其百分率的置信限可以根据文献[4]中的置信限等同性得到。

(3) 建立适用于正态分布截尾数据多元回归分析方法, 给出回归曲线的置信限和百分位值的置信区间估计。

(4) 能够全面开发利用不同条件下试验数据之间的横向信息, 在试样数相同的情况下, 比传统

方法具有更高的精度, 而在精度相同的情况下, 则可以节省大量试样。

(5) 定时截尾数据可参照文献[3]方法处理。

参考文献

- [1] Olive Jean Dunn, Virginia A. Clark. Applied Statistics: A Analysis of V ariance and Regression [M]. New York: John Wiley & Sons, Inc., 1987.
- [2] 傅惠民, 黄伟. 最佳无偏整体估计方法[J]. 机械强度, 2003, 25(3): 319- 324.
- [3] 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.
- [4] 傅惠民. 连续母体百分位值和百分率的区间估计和假设检验[J]. 航空学报, 1993, 14(11): 569- 577.