

文章编号: 1000-8055(2003)04-0465-05

截尾数据线性回归分析方法

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: 提出一种新的回归分析方法, 建立了截尾数据回归方程, 给出回归系数和标准差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最常见的正态分布截尾数据一元线性回归分析, 给出其回归曲线的置信限和百分位值的置信区间估计, 并将该方法推广到极值分布和威布尔分布的情况。传统回归分析只适用于来自正态分布的完全数据, 对于截尾数据或威布尔分布数据则无能为力, 目前工程上只能采用成组试验方法, 结果导致试验量很大。本文方法能够很好地解决上述问题, 与成组试验法相比, 在试样数相同的条件下具有更高的精度, 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; 回归分析; 最佳无偏整体估计; 截尾试验; 正态分布; 极值分布; 威布尔分布

中图分类号: TB301; O211 **文献标识码:** A

Linear Regression Analysis Method for Censored Data

HUANG Wei, FU Huimin

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

Abstract: A new regression analysis method is presented and the regression equation for censored data is established. The best unbiased integral estimators for the regression parameters, the standard deviation and their covariance are also given. The confidence limits of the regression curve and the confidence intervals of percentiles for normal distributions are discussed in detail. The traditional regression analysis is only suitable to complete data from normal distribution. In engineering, censored data is analyzed by group test method that requires a large number of samples. The present method extends traditional regression analysis to not only censored data but also extreme distribution and Weibull distribution. Compared with traditional method, the present method can reach a higher precision and a great number of specimens can be saved.

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

收稿日期: 2003-05-20; **修订日期:** 2003-07-01

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

回归分析是一种强有力的数据处理工具,在自然科学和社会科学的各个领域都有广泛的应用,但传统的回归分析只适用于来自正态分布的完全数据^[1],对于工程上常见的截尾数据和极值分布、威布尔分布的情况,则无能为力。目前对于截尾数据通常采用成组试验法,该方法对不同条件下的试验数据只能分别进行处理,这就要求在每种条件下都做大量试验,结果导致总的试验量很大。对此,文献[2]提出整体分布参数的概念,文献[3]建立了最佳无偏整体估计方法,本文进一步给出一种截尾数据回归分析方法,详细讨论了正态分布截尾数据的一元线性回归分析,并将其推广到极值分布和威布尔分布的情况。本文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,从而能够全面开发利用不同条件下试验数据之间的横向信息,使其可利用的信息量远远大于传统方法,既能够提高统计推断的精度,又可以节省大量试样。

2 正态分布截尾数据线性回归分析

正态分布是最常见同时也是最重要的分布,工程中许多随机变量均服从这种分布。设随机变量 y 与自变量 x 之间成线性关系:

$$y = a + bx + \epsilon \quad \epsilon \sim N(0, \sigma^2) \quad (1)$$

式中 a , b 和 σ 为待定参数。

设在状态 x_j 进行 n_j 个试样的试验, $y_{j1} \leq y_{j2} \leq \dots \leq y_{jq_j}$ 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[3]可知, a , b , σ 的最佳线性无偏估计量 $\hat{a}$, $\hat{b}$, $\hat{\sigma}$ 由下面诸式给出:

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

$$\hat{b} = \frac{L_{11}L_{22} - L_{21}L_{12}}{L_{11}L_{22} - L_{12}^2} \quad (3)$$

$$\hat{\sigma} = \frac{L_{22}L_{11} - L_{12}L_{21}}{L_{11}L_{22} - L_{12}^2} \quad (4)$$

回归方程为:

$$\hat{y} = \hat{a} + \hat{b}x \quad (5)$$

式中:

$$n' = \sum_{j=1}^m \sum_{k=1}^{q_j} g_{jkl} \quad (6)$$

$$\bar{x} = \frac{1}{n'} \sum_{j=1}^m \sum_{k=1}^{q_j} g_{jkl} x_j \quad (7)$$

$$\bar{u} = \frac{1}{n'} \sum_{j=1}^m \sum_{k=1}^{q_j} g_{jkl} u_{jk} \quad (8)$$

$$\bar{y} = \frac{1}{n'} \sum_{j=1}^m \sum_{k=1}^{q_j} g_{jkl} y_{jk} \quad (9)$$

$$L_{11} = \sum_{j=1}^m \sum_{k=1}^{q_j} g_{jkl} (x_j - \bar{x})(x_j - \bar{x}) \quad (10)$$

$$L_{12} = \sum_{j=1}^m \sum_{k=1}^{q_j} g_{jkl} (x_j - \bar{x})(u_{jk} - \bar{u}) \quad (11)$$

$$L_{22} = \sum_{j=1}^m \sum_{k=1}^{q_j} g_{jkl} (u_{jk} - \bar{u})(u_{jl} - \bar{u}) \quad (12)$$

$$L_{1y} = \sum_{j=1}^m \sum_{k=1}^{q_j} g_{jkl} (x_j - \bar{x})(y_{jk} - \bar{y}) \quad (13)$$

$$L_{2y} = \sum_{j=1}^m \sum_{k=1}^{q_j} g_{jkl} (u_{jl} - \bar{u})(y_{jk} - \bar{y}) \quad (14)$$

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

$$u_{jk} = \int \frac{n_j!}{(k-1)!(n_j-k)!} p_k^{k-1} \times (1-p_k)^{n_j-k} \Phi^{-1}(p_k) dp_k \quad (15)$$

$$v_{jkk} = \int \left\{ \frac{n_j!}{(k-1)!(n_j-k)!} p_k^{k-1} \times (1-p_k)^{n_j-k} [\Phi^{-1}(p_k)]^2 \right\} dp_k - u_{jk}^2 \quad (16)$$

$$v_{jkl} = \iint \left\{ \frac{n_j!}{(k-1)!(l-k-1)!(n_j-l)!} \times p_k^{k-1} (p_l - p_k)^{l-k-1} (1-p_l)^{n_j-l} \Phi^{-1}(p_k) \times \Phi^{-1}(p_l) \right\} dp_k dp_l - u_{jk} u_{jl} \quad (k < l) \quad (17)$$

可以证明, $\hat{a}$, $\hat{b}$, $\hat{\sigma}$ 是 a , b , σ 的最佳线性无偏估计。大量 Monte Carlo 模拟表明,当数据总量 $q =$

$\sum_{j=1}^m q_j$ 稍大时,即可近似为正态分布。

3 均值和标准差的置信限估计

均值 $\mu(y) = a + bx$ 和标准差 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:

$$\mu_L(x) = \hat{a} + \hat{b}x - \frac{\hat{u}_\gamma \hat{\sigma} \sqrt{c_{11} + 2c_{12}x + c_{22}x^2}}{\hat{u}_\gamma \hat{\sigma} \sqrt{c_{11} + 2c_{12}x + c_{22}x^2}} \quad (18)$$

$$\mu_U(x) = \hat{a} + \hat{b}x + \frac{\hat{u}_\gamma \hat{\sigma} \sqrt{c_{11} + 2c_{12}x + c_{22}x^2}}{\hat{u}_\gamma \hat{\sigma} \sqrt{c_{11} + 2c_{12}x + c_{22}x^2}} \quad (19)$$

$$\sigma_L = (1 - u_\gamma \sqrt{c_{33}}) \hat{\sigma} \quad (20)$$

$$\sigma_U = (1 + u_\gamma \sqrt{c_{33}}) \hat{\sigma} \quad (21)$$

式中:

$$(C_{kl})_{3 \times 3} =$$
$$\begin{bmatrix} \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} & \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} X_j & \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} U_{jk} \\ \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} X_j & \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} X_j^2 & \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} U_{jk} X_j \\ \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} U_{jk} & \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} U_{jk} X_j & \sum_{j=1}^m \sum_{k,l=1}^{q_j} g_{jkl} U_{jk} U_{jl} \end{bmatrix}^{-1}$$

(22)

4 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 $y_P(x) = a + bx + u_P \sigma$ 的无偏估计量 $\hat{y}_P(x)$ 由下式给出:

$$\hat{y}_P(x) = \hat{a} + \hat{b}x + \hat{u}_P \hat{\sigma} \tag{23}$$

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

$$y_{PL}(x) = \hat{a} + \hat{b}x + \hat{u}_P \hat{\sigma} -$$
$$u_P \left\{ C_{11} + 2C_{13}u_P + C_{33}u_P^2 + \right.$$
$$\left. 2(C_{12} + C_{23}u_P)x + C_{22}x^2 \right\}^{1/2}$$

(24)

$$y_{PU}(x) = \hat{a} + \hat{b}x + \hat{u}_P \hat{\sigma} +$$
$$u_P \left\{ C_{11} + 2C_{13}u_P + C_{33}u_P^2 + \right.$$
$$\left. 2(C_{12} + C_{23}u_P)x + C_{22}x^2 \right\}^{1/2}$$

(25)

此外, 根据百分位值的置信度为 γ 的单侧置信上(下)限曲线与百分率的置信度为 γ 的单侧置信下(上)限曲线相同的原理^[4], 可以对随机变量 y 的百分率置信限进行估计。

5 定时截尾数据的处理

定数截尾试验数据可以直接采用上述方法处理。但是对定时截尾试验数据, 如果直接采用上述方法, 则将损失最后一个失效数据到试验截止时间 τ 这一时间段内的试验信息。下面给出一种能够利用该试验信息的方法。设在状态 x_j 下进行定时截尾试验, $y_{j1} \leq y_{j2} \leq \dots \leq y_{jq_j} \leq \tau_j$, 为得到的前 q_j 个定时截尾样本数据。如果将截尾时间 τ_j 也看成一个有序数据, 则可以将定时截尾数据等效成定数截尾数据。由于截尾时间 τ_j 位于 y_{jq_j} 和下一个失效数据 $y_{j(q_j+1)}$ 之间, 即 $y_{jq_j} \leq \tau_j \leq y_{j(q_j+1)}$, 所以与有序数据 τ_j 相对应的 $u_{j(q_j+1)}^*$ 、 $v_{jk(q_j+1)}^*$ 和 $v_{j(q_j+1)(q_j+1)}^*$ 由下面诸式计算:

$$u_{j(q_j+1)}^* = (u_{jq_j} + u_{j(q_j+1)})/2$$

(26)

$$v_{jk(q_j+1)}^* = (v_{jkq_j} + v_{jk(q_j+1)})/2$$
$$k = 1, 2, \dots, q_j$$

(27)

$$v_{j(q_j+1)(q_j+1)}^* = (v_{jq_jq_j} + v_{j(q_j+1)(q_j+1)})/2$$

(28)

6 极值分布和威布尔分布一元回归分析

对于随机变量 y 遵循极值分布 $EV(\mu, \sigma)$ 的情况, 设

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

(29)

只要将式(15)、(16)和(17)改成下面三式:

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

(30)

$$v_{jkk} = \int \left\{ \frac{n_i!}{(k-1)!(n_j-k)!} p_k^{k-1} (1-p_k)^{n_j-k} \times \right.$$
$$\left. \left[\ln \ln \frac{1}{1-p_k} \right]^2 \right\} dp_k - u_{jk}^2$$

(31)

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

(32)

并将式(23)、(24)和(25)改换成:

$$\hat{y}_P(x) = \hat{a} + \hat{b}x + \hat{\sigma} \ln \ln \frac{1}{1-P}$$

(33)

$$y_{PL}(x) = \hat{a} + \hat{b}x + \hat{\sigma} \ln \ln \frac{1}{1-P} - u_P \left\{ C_{11} + \right.$$
$$2C_{13} \ln \ln \frac{1}{1-P} + C_{33} \left[\ln \ln \frac{1}{1-P} \right]^2 +$$
$$\left. 2 \left[C_{12} + C_{23} \ln \ln \frac{1}{1-P} \right] x + C_{22} x^2 \right\}^{1/2}$$

(34)

$$y_{PU}(x) = \hat{a} + \hat{b}x + \hat{\sigma} \ln \ln \frac{1}{1-P} + u_P \left\{ C_{11} + \right.$$
$$2C_{13} \ln \ln \frac{1}{1-P} + C_{33} \left[\ln \ln \frac{1}{1-P} \right]^2 +$$
$$\left. 2 \left[C_{12} + C_{23} \ln \ln \frac{1}{1-P} \right] x + C_{22} x^2 \right\}^{1/2}$$

(35)

则本文方法和公式对极值分布也同样成立。

对于随机变量 t 遵循威布尔分布 $W(\alpha, \beta)$ 的情况, 令 $y = \ln t$, 则 y 遵循极值分布 $EV(\mu, \sigma)$, 可按上述方法处理, 其中 $\mu = \ln \beta$, $\sigma = 1/\alpha$ 。

7 应用举例

例 1: 表 1 给出 5 组不同应力水平下的定数截尾寿命数据, 每组 5 个试样。它们服从对数正态分布 $LN(\mu, \sigma^2)$, 且方差 σ^2 为常数。根据逆幂律模型, 均值 μ 与对数应力水平 $\ln S$ 之间成线性关系:

$\mu(S) = a + b\ln S$

表 1 定数截尾寿命数据与回归分析结果

Table 1 Censored life data and regression analysis results

序 号	1	2	3	4	5
应力/MPa	750	800	850	900	950
寿命/ 10^3c	462 1	429 2	195 5	86 3	69 1
		458 1	245 9	111 3	72 8
			259 4	121 3	78 2
					96 9
					118 1
t_{50}	667.3	379.5	219.8	127.5	74.1
t_{90}	495.2	278.4	158.1	90.1	52.0
t_{99}	370.6	204.3	114.3	65.0	37.6

根据本文方法对表 1 中数据进行回归分析, 得到参数 a 、 b 和 σ 的最佳无偏整体估计:

$\hat{a} = 76\ 6421, \hat{b} = -9\ 5247, \hat{\sigma} = 0.2487$

其协方差矩阵为:

$$\text{Cov} \begin{bmatrix} \hat{a} \\ \hat{b} \\ \hat{\sigma} \end{bmatrix} = \begin{bmatrix} 27.2985 & -4.0295 & 0.1178 \\ -4.0295 & 0.5949 & -0.01716 \\ 0.1178 & -0.01716 & 0.004005 \end{bmatrix}$$

根据式(18)、(19)、(20)和(21)求得均值 μ 和标准差 σ 的置信上、下限公式:

$$\begin{aligned} \mu_U(S) &= 76\ 6421 - 9\ 5247 \ln S + \\ &\quad u_Y \sqrt{27.2985 - 8.0590 \ln S + 0.5949 \ln^2 S} \\ \mu_L(S) &= 76\ 6421 - 9\ 5247 \ln S - \\ &\quad u_Y \sqrt{27.2985 - 8.0590 \ln S + 0.5949 \ln^2 S} \\ \sigma_U &= 0.2487 + 0.06329 u_Y \\ \sigma_L &= 0.2487 - 0.06329 u_Y \end{aligned}$$

再由式(24)得到置信度为 Y 、概率为 P 的对数安全寿命置信下限:

$$\begin{aligned} y_{PL}(S) &= 76\ 6421 - 9\ 5247 \ln S + 0.2487 u_P - \\ &\quad u_Y \{ 27.2985 - 8.0590 \ln S + 0.5949 \ln^2 S + \\ &\quad 0.004005 u_P^2 + 0.2356 u_P - 0.03432 u_P \ln S \}^{1/2} \end{aligned}$$

当取 $Y = 0.9$ 和 $P = 0.5, 0.1, 0.01$ 时, 由上式即可求得置信度为 90% 和可靠度分别为 50%、90%、99% 的安全寿命置信下限 t_{50}, t_{90}, t_{99} , 结果列于表 1 中。

例 2: 表 2 给出某产品四组不同应力水平下的定时截尾寿命数据^[5], 它们服从对数正态分布 $LN(\mu, \sigma^2)$, 且方差 σ^2 为常数。每个温度均进行了

25 个产品的定时截尾试验, 其截尾时间分别为:

$T_1 = 370\text{K}, \tau_1 = 7000\text{h}; T_2 = 413\text{K}, \tau_2 = 1800\text{h};$
 $T_3 = 440\text{K}, \tau_3 = 1100\text{h}; T_4 = 463\text{K}, \tau_4 = 500\text{h}.$

根据阿伦尼斯模型, 对数寿命均值 μ 与温度 T 之间成线性关系:

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

表 2 定时截尾寿命数据与回归分析结果

Table 2 Censored life data and regression analysis results

序 号	1	2	3	4
温度/K	370	413	440	463
寿命/h	529	265	66	28
	822	336	149	41
	1905	449	158	56
	2102	563	162	59
	2756	582	171	59
	3612	627	188	64
	4617	766	194	74
	5275	825	244	94
	5446	873	312	101
	6016	924	335	108
	6645	924	342	148
	6859	956	387	156
		1102	409	156
		1116	543	162
		1148	581	164
		1209	664	165
		1487	667	246
			789	264
			1064	289
				377
				443
				490
t_{50}	5716.43	912.27	321.14	140.23
t_{90}	1690.88	257.71	89.56	39.17
t_{99}	571.67	83.76	29.17	12.88

采用本文方法对表 2 中试验数据进行处理, 得到参数 a 、 b 和 σ 的最佳无偏整体估计:

$\hat{a} = -9\ 8297, \hat{b} = 6934.29, \hat{\sigma} = 0.9413$

其协方差矩阵为:

$$\text{Cov} \begin{bmatrix} \hat{a} \\ \hat{b} \\ \hat{\sigma} \end{bmatrix} = \begin{bmatrix} 1.5732 & -656.48 & -0.02316 \\ -656.48 & 276077.7 & 11.6485 \\ -0.02316 & 11.6485 & 0.01522 \end{bmatrix}$$

根据式(18)、(19)、(20)和(21)求得均值 μ 和

标准差 σ 的置信上、下限公式:

$$\begin{aligned}\mu_U(T) &= -9\,8297 + 6934\,29/T + \\ &\quad u_Y \sqrt{1\,5732 - 1312\,96/T + 276077\,7/T^2} \\ \mu_L(T) &= -9\,8297 + 6934\,29/T - \\ &\quad u_Y \sqrt{1\,5732 - 1312\,96/T + 276077\,7/T^2} \\ \alpha_U &= 0\,9413 + 0\,1234u_Y \\ \alpha_L &= 0\,9413 - 0\,1234u_Y\end{aligned}$$

再由式(24)得到置信度为 γ 概率为 P 的对数安全寿命置信下限:

$$\begin{aligned}y_{PL}(T) &= -9\,8297 + 6934\,29/T + 0\,9413u_P - \\ &\quad u_Y \{1\,5732 - 1312\,96/T + 276077\,7/T^2 - \\ &\quad 0\,04632u_P + 0\,01522u_P^2 + 23\,2970u_P/T\}^{1/2}\end{aligned}$$

当取 $\gamma = 0\,9$ 和 $P = 0\,5, 0\,1, 0\,01$ 时, 由上式即可求得置信度为 90% 和可靠度分别为 50%、90%、99% 的安全寿命置信下限 $t_{50\%}, t_{90\%}, t_{99\%}$ 列于表 2 中。

8 结 论

- (1) 提出一种截尾数据回归分析方法, 将只适用于完全数据的传统回归分析推广到工程中常见的截尾数据情况。
- (2) 详细讨论了工程中最常见的正态分布截尾数据的一元线性回归分析, 给出回归系数和标准差的最佳无偏整体估计及其协方差矩阵。建立回归曲线的置信限和百分位值的置信区间估计。
- (3) 建立适用于威布尔分布、极值分布的一元线性回归分析方法。

(4) 给出一种将定时截尾数据等效成定数截尾试验数据的方法。

(5) 本文方法对一种条件下只有一个失效数据的情况也能很好地进行处理, 这一特性允许加速寿命试验中选取的最小加速应力尽量接近正常使用情况, 从而保证将加速试验结果外推到正常使用情况时有较高的精度。

(6) 本文方法能够全面开发利用不同条件下试验数据之间的横向信息, 使其可利用的信息量远远大于只能分别在各自条件下对试验数据进行统计分析的成组试验方法。工程应用表明, 在试样数相同的情况下, 本文方法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。

参考文献:

[1] Olive Jean Dunn, Virginia A. Clark Applied Statistics: A nalysis of V ariance and Regression [M], New York: John Wiley & Sons, Inc, 1987

[2] 傅惠民 整体推断的极大似然方法[J] 机械强度, 2002, 24 (1): 1- 5

[3] 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

[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.

[5] 茆诗松, 王玲玲 加速寿命试验[M]. 北京: 科学出版社, 1997.