

文章编号: 1000-8055(2004)01-0001-06

截尾数据异方差回归分析

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: 提出截尾数据异方差回归分析方法, 将传统的只适用于同方差和完全数据的回归分析推广到异方差和截尾数据的情况。文中建立了异方差回归方程, 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。详细讨论了正态分布、极值分布、威布尔分布情况下的线性异方差回归分析, 给出其回归曲线和百分位值的置信限。对于异方差问题目前工程上只能采用成组试验方法, 因此试验量很大; 本文方法与之相比则能很好地解决这一问题, 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, 可以节省大量试样。

关键词: 航空、航天推进系统; 异方差回归分析; 截尾数据; 最佳无偏整体估计; 可靠性

中图分类号: TB114; O211

文献标识码: A

Heteroscedasticity Regression Analysis for Censored Data

HUANG Wei, FU Huimin

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

Abstract: A heteroscedasticity regression analysis method for censored data is presented, and it extends the traditional regression analysis, which is only suitable to complete and homoscedasticity data, to censored and heteroscedastic data. The heteroscedasticity regression equation is established and the best linear unbiased estimators for the regression parameters, the standard deviation and their covariance are also given. The linear heteroscedasticity regression analysis methods for normal distribution, extreme distribution and Weibull distribution are discussed in detail and the confidence limits of the regression curve and percentiles are given. At present, heteroscedastic data can only be analyzed by group test method that requires large sample. Compared with traditional method, the present method can reach higher precision and a great number of specimens can be saved.

Key words: aerospace propulsion system; heteroscedasticity regression analysis; censored data; best unbiased integral estimate; reliability

传统的回归分析只适用于来自正态分布的同方差完全数据^[1], 但工程上常常遇到异方差问题, 目前只能采用成组试验法对异方差数据进行处理。这种方法不能将所有条件下的数据作为一个

整体进行统计分析, 从而要求在每种条件下都做大量试验, 导致总的试验量很大。文献[2]提出了最佳无偏整体估计的一般原理和方法, 文献[3, 4]给出了同方差截尾数据的回归分析, 本文进一步

收稿日期: 2003-10-18; **修订日期:** 2004-1-12

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

建立异方差回归分析方法。

1 一元异方差回归分析

1.1 回归系数估计

设正态随机变量 y 与自变量 x 成线性关系:

$$y = a_0 + a_1 x + \epsilon$$

$$\epsilon \sim N(0, \sigma^2(x)) \quad (1)$$

其中:

$$\sigma(x) = b_0 + b_1 x \quad (2)$$

式中 a_0, a_1 和 b_0, b_1 为待定参数。设在状态 x_j 下进行 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 个样本数据, $j = 1, 2, \dots, m$ 。记 $\sum_{j,k,l} =$

$\sum_{j=1}^m \sum_{k=1}^{q_j}$, 则根据文献[2]可知, a_0, a_1 和 b_0, b_1 的估计量 $\hat{a}_0, \hat{a}_1$ 和 $\hat{b}_0, \hat{b}_1$ 由下面诸式给出:

$$\hat{a}_0 = \bar{y} - \hat{a}_1 \bar{x} - \hat{b}_0 \bar{u} - \hat{b}_1 \bar{xu} \quad (3)$$

$$\hat{\theta} = (\hat{a}_1, \hat{b}_0, \hat{b}_1)^T = L_{xx}^{-1} L_{xy} \quad (4)$$

式中:

$$n' = \sum_{j,k,l} \frac{g_{jkl}}{\sigma_j^2} \quad (5)$$

$$\bar{x} = \frac{1}{n'} \sum_{j,k,l} \frac{g_{jkl} x_j}{\sigma_j^2} \quad (6)$$

$$\bar{u} = \frac{1}{n'} \sum_{j,k,l} \frac{g_{jkl} u_{jk}}{\sigma_j^2} \quad (7)$$

$$\bar{y} = \frac{1}{n'} \sum_{j,k,l} \frac{g_{jkl} y_{jk}}{\sigma_j^2} \quad (8)$$

$$\bar{xu} = \frac{1}{n'} \sum_{j,k,l} \frac{g_{jkl} x_j u_{jk}}{\sigma_j^2} \quad (9)$$

$$L_{xx} = (l_{11} l_{22})_{3 \times 3} \quad (10)$$

$$l_{11} = \sum_{j,k,l} \frac{g_{jkl} (x_j - \bar{x})^2}{\sigma_j^2} \quad (11)$$

式中:

$$C = \begin{bmatrix} \sum_{j,k,l} \frac{g_{jkl}}{\sigma_j^2} & \sum_{j,k,l} \frac{g_{jkl} x_j}{\sigma_j^2} & \sum_{j,k,l} \frac{g_{jkl} u_{jk}}{\sigma_j^2} & \sum_{j,k,l} \frac{g_{jkl} x_j u_{jk}}{\sigma_j^2} \\ \sum_{j,k,l} \frac{g_{jkl} x_j}{\sigma_j^2} & \sum_{j,k,l} \frac{g_{jkl} x_j^2}{\sigma_j^2} & \sum_{j,k,l} \frac{g_{jkl} x_j u_{jk}}{\sigma_j^2} & \sum_{j,k,l} \frac{g_{jkl} x_j^2 u_{jk}}{\sigma_j^2} \\ \sum_{j,k,l} \frac{g_{jkl} u_{jk}}{\sigma_j^2} & \sum_{j,k,l} \frac{g_{jkl} x_j u_{jk}}{\sigma_j^2} & \sum_{j,k,l} \frac{g_{jkl} u_{jk}^2}{\sigma_j^2} & \sum_{j,k,l} \frac{g_{jkl} x_j u_{jk}^2}{\sigma_j^2} \\ \sum_{j,k,l} \frac{g_{jkl} x_j u_{jk}}{\sigma_j^2} & \sum_{j,k,l} \frac{g_{jkl} x_j^2 u_{jk}}{\sigma_j^2} & \sum_{j,k,l} \frac{g_{jkl} x_j u_{jk}^2}{\sigma_j^2} & \sum_{j,k,l} \frac{g_{jkl} x_j^2 u_{jk}^2}{\sigma_j^2} \end{bmatrix}^{-1} \quad (22)$$

如果不进行迭代计算, 而直接在式(5)~(20)中取 $\sigma_j^2 = 1$, 则式(3)和(4)给出 a_0, a_1 和 b_0, b_1 的无偏估计量, 且它们的协方差矩阵为:

$$\text{Cov}(\hat{a}_0, \hat{a}_1, \hat{b}_0, \hat{b}_1)^T = C = E^{-1} D E^{-1} \quad (23)$$

$$l_{12} = l_{21} = \sum_{j,k,l} \frac{g_{jkl} (x_j - \bar{x})(u_{jk} - \bar{u})}{\sigma_j^2} \quad (12)$$

$$l_{13} = l_{31} = \sum_{j,k,l} \frac{g_{jkl} (x_j u_{jk} - \bar{x} \bar{u})(x_j - \bar{x})}{\sigma_j^2} \quad (13)$$

$$l_{22} = \sum_{j,k,l} \frac{g_{jkl} (u_{jk} - \bar{u})(u_{jk} - \bar{u})}{\sigma_j^2} \quad (14)$$

$$l_{23} = l_{32} = \sum_{j,k,l} \frac{g_{jkl} (x_j u_{jk} - \bar{x} \bar{u})(u_{jk} - \bar{u})}{\sigma_j^2} \quad (15)$$

$$l_{33} = \sum_{j,k,l} \frac{g_{jkl} (x_j u_{jk} - \bar{x} \bar{u})(x_j u_{jk} - \bar{x} \bar{u})}{\sigma_j^2} \quad (16)$$

$$L_{xy} = (l_{1y}, l_{2y}, l_{3y})^T \quad (17)$$

$$l_{1y} = \sum_{j,k,l} \frac{g_{jkl} (x_j - \bar{x})(y_{jk} - \bar{y})}{\sigma_j^2} \quad (18)$$

$$l_{2y} = \sum_{j,k,l} \frac{g_{jkl} (u_{jk} - \bar{u})(y_{jk} - \bar{y})}{\sigma_j^2} \quad (19)$$

$$l_{3y} = \sum_{j,k,l} \frac{g_{jkl} (x_j u_{jk} - \bar{x} \bar{u})(y_{jk} - \bar{y})}{\sigma_j^2} \quad (20)$$

其中 $\sigma_j^2 = \sigma(x_j)$; $(g_{jkl})_{q_j \times q_j} = (v_{jkl})_{q_j \times q_j}^{-1}$, 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^[3]。

当 σ_j^2 已知时, 可以证明 $\hat{a}_0, \hat{a}_1$ 和 $\hat{b}_0, \hat{b}_1$ 是 a_0, a_1 和 b_0, b_1 的最佳无偏估计; 当 σ_j^2 未知时, 可先取 $\sigma_j^2 = 1$ 代入(5)~(20)求得 $\hat{a}_0, \hat{a}_1$ 和 $\hat{b}_0, \hat{b}_1$ 的初值, 并计算 σ_j^2 的估计值 $\hat{\sigma}_j^2 = b_0 + b_1 x_j$, 然后反复迭代计算, 直到 $\hat{\sigma}_j^2$ 稳定为止。大量 Monte Carlo 模拟表明, 当数据总量稍大时, $\hat{a}_0, \hat{a}_1$ 和 $\hat{b}_0, \hat{b}_1$ 即可用正态分布近似, 它们的协方差矩阵为:

$$\text{Cov}(\hat{a}_0, \hat{a}_1, \hat{b}_0, \hat{b}_1)^T = C \quad (21)$$

$$E = \begin{bmatrix} \sum_{j,k,l} g_{jkl} & \sum_{j,k,l} g_{jkl} X_j & \sum_{j,k,l} g_{jkl} u_{jk} & \sum_{j,k,l} g_{jkl} X_j u_{jk} \\ \sum_{j,k,l} g_{jkl} X_j & \sum_{j,k,l} g_{jkl} X_j^2 & \sum_{j,k,l} g_{jkl} X_j u_{jk} & \sum_{j,k,l} g_{jkl} X_j^2 u_{jk} \\ \sum_{j,k,l} g_{jkl} u_{jk} & \sum_{j,k,l} g_{jkl} X_j u_{jk} & \sum_{j,k,l} g_{jkl} u_{jk} u_{jl} & \sum_{j,k,l} g_{jkl} X_j u_{jk} u_{jl} \\ \sum_{j,k,l} g_{jkl} X_j u_{jk} & \sum_{j,k,l} g_{jkl} X_j^2 u_{jk} & \sum_{j,k,l} g_{jkl} X_j u_{jk} u_{jl} & \sum_{j,k,l} g_{jkl} X_j^2 u_{jk} u_{jl} \end{bmatrix} \quad (24)$$

$$D = \begin{bmatrix} \sum_{j,k,l} \hat{\sigma}_j^2 g_{jkl} & \sum_{j,k,l} \hat{\sigma}_j^2 g_{jkl} X_j & \sum_{j,k,l} \hat{\sigma}_j^2 g_{jkl} u_{jk} & \sum_{j,k,l} \hat{\sigma}_j^2 g_{jkl} X_j u_{jk} \\ \sum_{j,k,l} \hat{\sigma}_j^2 g_{jkl} X_j & \sum_{j,k,l} \hat{\sigma}_j^2 g_{jkl} X_j^2 & \sum_{j,k,l} \hat{\sigma}_j^2 g_{jkl} X_j u_{jk} & \sum_{j,k,l} \hat{\sigma}_j^2 g_{jkl} X_j^2 u_{jk} \\ \sum_{j,k,l} \hat{\sigma}_j^2 g_{jkl} u_{jk} & \sum_{j,k,l} \hat{\sigma}_j^2 g_{jkl} X_j u_{jk} & \sum_{j,k,l} \hat{\sigma}_j^2 g_{jkl} u_{jk} u_{jl} & \sum_{j,k,l} \hat{\sigma}_j^2 g_{jkl} X_j u_{jk} u_{jl} \\ \sum_{j,k,l} \hat{\sigma}_j^2 g_{jkl} X_j u_{jk} & \sum_{j,k,l} \hat{\sigma}_j^2 g_{jkl} X_j^2 u_{jk} & \sum_{j,k,l} \hat{\sigma}_j^2 g_{jkl} X_j u_{jk} u_{jl} & \sum_{j,k,l} \hat{\sigma}_j^2 g_{jkl} X_j^2 u_{jk} u_{jl} \end{bmatrix} \quad (25)$$

1.2 回归方程及百分位值曲线估计

随机变量 y 的回归方程为:

$$\hat{y}(x) = \hat{a}_0 + \hat{a}_1 x \quad (26)$$

回归曲线的置信度为 γ 的单侧置信下限和单侧置信上限分别由下面两式给出:

$$y_L(x) = \hat{a}_0 + \hat{a}_1 x - u_\gamma \sqrt{c_{11} + 2c_{12}x + c_{22}x^2} \quad (27)$$

$$y_U(x) = \hat{a}_0 + \hat{a}_1 x + u_\gamma \sqrt{c_{11} + 2c_{12}x + c_{22}x^2} \quad (28)$$

式中 $c_{i_1 i_2}$ 为协方差矩阵 C 的元素。

标准差 $\sigma(x)$ 的点估计为:

$$\hat{\sigma}(x) = \hat{b}_0 + \hat{b}_1 x \quad (29)$$

$\sigma(x)$ 的置信度为 γ 的单侧置信下限和单侧置信上限分别为:

$$\alpha_L(x) = \hat{b}_0 + \hat{b}_1 x - u_\gamma \sqrt{c_{33} + 2c_{34}x + c_{44}x^2} \quad (30)$$

$$\alpha_U(x) = \hat{b}_0 + \hat{b}_1 x + u_\gamma \sqrt{c_{33} + 2c_{34}x + c_{44}x^2} \quad (31)$$

随机变量 y 的概率为 P 的百分位值 $y_P(x)$ 的估计量由下式给出

$$y_P(x) = \hat{a}_0 + u_P \hat{b}_0 + (\hat{a}_1 + u_P \hat{b}_1)x \quad (32)$$

$y_P(x)$ 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:

$$y_{PL}(x) = \hat{a}_0 + u_P \hat{b}_0 + (\hat{a}_1 + u_P \hat{b}_1)x -$$

$$u_\gamma \sqrt{c_{11} + 2u_P c_{13} + u_P^2 c_{33} + 2(c_{12} + u_P c_{14} + u_P c_{23} + u_P c_{34})x + (c_{22} + 2u_P c_{24} + u_P^2 c_{44})x^2} \quad (33)$$

$$y_{PU}(x) = \hat{a}_0 + u_P \hat{b}_0 + (\hat{a}_1 + u_P \hat{b}_1)x +$$

$$u_\gamma \sqrt{c_{11} + 2u_P c_{13} + u_P^2 c_{33} + 2(c_{12} + u_P c_{14} + u_P c_{23} + u_P c_{34})x + (c_{22} + 2u_P c_{24} + u_P^2 c_{44})x^2} \quad (34)$$

2 极值分布和威布尔分布异方差回归分析

设随机变量 y 遵循极值分布 $EV(\mu, \sigma)$, y 与 x 之间成线性关系:

$$y = a_0 + a_1 x + \epsilon \quad \epsilon \sim EV(0, \sigma(x)) \quad (35)$$

其中:

$$\sigma(x) = b_0 + b_1 x \quad (36)$$

此时, 只要将第二节中的 u_{jk} 和 v_{jkl} 改为标准极值分布顺序统计量的均值和协方差, 则式(3)~(25)对极值分布仍然成立。其回归方程为:

$$\hat{y}(x) = \hat{a}_0 - \omega \hat{b}_0 + (\hat{a}_1 - \omega \hat{b}_1)x \quad (37)$$

式中 $\omega \approx 0.5772$ 为欧拉常数, 回归曲线的置信度为 γ 的单侧置信下限 $y_L(x)$ 和单侧置信上限 $y_U(x)$ 为:

$$y_L(x) = \hat{a}_0 - \omega \hat{b}_0 + (\hat{a}_1 - \omega \hat{b}_1)x -$$

$$u_Y \sqrt{c_{11} - 2\omega c_{13} + \omega^2 c_{33} + 2(c_{12} - \omega c_{14} - \omega c_{23} + \omega^2 c_{34})x + (c_{22} - 2\omega c_{24} + \omega^2 c_{44})x^2}$$

$$y_U(x) = \hat{a}_0 - \omega \hat{b}_0 + (\hat{a}_1 - \omega \hat{b}_1)x +$$

$$u_Y \sqrt{c_{11} - 2\omega c_{13} + \omega^2 c_{33} + 2(c_{12} - \omega c_{14} - \omega c_{23} + \omega^2 c_{34})x + (c_{22} - 2\omega c_{24} + \omega^2 c_{44})x^2} \quad (39)$$

随机变量 y 的概率为 P 的百分位值 $y_P(x)$ 的无偏估计量由下式给出

$$\hat{y}_P(x) = \hat{a}_0 + \hat{b}_0 \ln \ln \frac{1}{1-P} + \left[\hat{a}_1 + \hat{b}_1 \ln \ln \frac{1}{1-P} \right] x \quad (40)$$

$y_P(x)$ 的置信度为 γ 的单侧置信下限 $y_{PL}(x)$ 和单侧置信上限 $y_{PU}(x)$ 分别为:

$$y_{PL}(x) = \hat{a}_0 + \hat{b}_0 \ln \ln \frac{1}{1-P} + \left[\hat{a}_1 + \hat{b}_1 \ln \ln \frac{1}{1-P} \right] x - u \left\{ c_{11} + 2c_{13} \ln \ln \frac{1}{1-P} + \right.$$

$$c_{33} \ln^2 \ln \frac{1}{1-P} + 2 \left[c_{12} + c_{14} \ln \ln \frac{1}{1-P} + c_{23} \ln \ln \frac{1}{1-P} + c_{34} \ln^2 \ln \frac{1}{1-P} \right] x +$$

$$\left. \left[c_{22} + 2c_{24} \ln \ln \frac{1}{1-P} + c_{44} \ln^2 \ln \frac{1}{1-P} \right] x^2 \right\}^{1/2} \quad (41)$$

$$y_{PU}(x) = \hat{a}_0 + \hat{b}_0 \ln \ln \frac{1}{1-P} + \left[\hat{a}_1 + \hat{b}_1 \ln \ln \frac{1}{1-P} \right] x + u \left\{ c_{11} + 2c_{13} \ln \ln \frac{1}{1-P} + \right.$$

$$c_{33} \ln^2 \ln \frac{1}{1-P} + 2 \left[c_{12} + c_{14} \ln \ln \frac{1}{1-P} + c_{23} \ln \ln \frac{1}{1-P} + c_{34} \ln^2 \ln \frac{1}{1-P} \right] x +$$

$$\left. \left[c_{22} + 2c_{24} \ln \ln \frac{1}{1-P} + c_{44} \ln^2 \ln \frac{1}{1-P} \right] x^2 \right\}^{1/2} \quad (42)$$

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 $EV(\mu, \sigma)$, 因此可按上述方法处理, 其中 $\mu = \ln \beta$, $\sigma = 1/\alpha$

$$\hat{a}_0 = \bar{y} - \hat{b}_0 \bar{u} - \sum_{i=1}^r \hat{a}_i \bar{x}_i - \sum_{i=1}^s \hat{b}_i \bar{x}_i \bar{u} \quad (45)$$

$$\hat{\theta} = (\hat{a}_1, \hat{a}_2, \dots, \hat{a}_r, \hat{b}_0, \hat{b}_1, \dots, \hat{b}_s)^T$$

$$= L_{xx}^{-1} L_{xy} \quad (46)$$

3 多元异方差回归分析方法

3.1 多元回归系数估计

设正态随机变量 y 与自变量 $x = (x_1, x_2, \dots, x_r)$ 之间成线性关系:

$$y = a_0 + \sum_{i=1}^r a_i x_i + \epsilon$$

$$\epsilon \sim N(0, \sigma^2(x)) \quad (43)$$

其中: $\sigma(x) = b_0 + \sum_{i=1}^s b_i x_i \quad (44)$

式中 $a_0, a_1, \dots, a_r$ 和 $b_0, b_1, \dots, b_s$ 为待定参数。设在状态 $x_j = (x_{j1}, x_{j2}, \dots, x_{jr})$ 下进行 n_j 个试样的试验, $y_{j1} \leq y_{j2} \leq \dots \leq y_{j n_j}$ 为观测到的前 q_j 个样本数据, 则根据文献[2]可知, $\hat{a}_0, \hat{a}_1, \dots, \hat{a}_r$ 和 $\hat{b}_0, \hat{b}_1, \dots, \hat{b}_s$ 的估计量 $\hat{a}_0, \hat{a}_1, \dots, \hat{a}_r$ 和 $\hat{b}_0, \hat{b}_1, \dots, \hat{b}_s$ 由下面诸式给出:

式中:

$$n' = \sum_{j,k,l} \frac{g_{jkl}}{\sigma_j^2} \quad (47)$$

$$\bar{x}_i = \frac{1}{n'} \sum_{j,k,l} \frac{g_{jkl} x_{ji}}{\sigma_j^2}$$

$$i = 1, 2, \dots, r \quad (48)$$

$$\bar{u} = \frac{1}{n'} \sum_{j,k,l} \frac{g_{jkl} u_{jk}}{\sigma_j^2} \quad (49)$$

$$\bar{y} = \frac{1}{n'} \sum_{j,k,l} \frac{g_{jkl} y_{jk}}{\sigma_j^2} \quad (50)$$

$$\bar{x}_i u = \frac{1}{n'} \sum_{j,k,l} \frac{g_{jkl} x_{ji} u_{jk}}{\sigma_j^2}$$

$$i = 1, 2, \dots, s \quad (51)$$

$$L_{xy} = (l_{1y}, l_{2y}, \dots, l_{(r+s+1)y})^T \quad (52)$$

$$L_{xx} = (l_{11}, l_{12}, \dots, l_{(r+s+1)(r+s+1)}) \quad (53)$$

$$l_{iy} = \begin{cases} \sum_{j,k,l} \frac{g_{jkl} (x_{ji} - \bar{x}_i) (y_{jk} - \bar{y})}{\sigma_j^2} & 1 \leq i \leq r \\ \sum_{j,k,l} \frac{g_{jkl} (u_{jl} - \bar{u}) (y_{jk} - \bar{y})}{\sigma_j^2} & i = r+1 \\ \sum_{j,k,l} \frac{g_{jkl} (x_{ji} u_{jk} - \bar{x}_i \bar{u}) (y_{jk} - \bar{y})}{\sigma_j^2} & r+2 \leq i \leq r+s+1 \end{cases} \quad (54)$$

$$l_{i_1 i_2} = \begin{cases} \sum_{j, k, l} \frac{g_{jkl} (x_{ji_1} - \bar{x}_{i_1}) (x_{ji_2} - \bar{x}_{i_2})}{\sigma_j^2} & 1 \leq i_1 \leq r, \quad 1 \leq i_2 \leq r \\ \sum_{j, k, l} \frac{g_{jkl} (x_{ji_1} - \bar{x}_{i_1}) (u_{jk} - \bar{u})}{\sigma_j^2} & 1 \leq i_1 \leq r, \quad i_2 = r+1 \\ \sum_{j, k, l} \frac{g_{jkl} (u_{jl} - \bar{u}) (u_{jk} - \bar{u})}{\sigma_j^2} & i_1 = r+1, \quad i_2 = r+1 \\ \sum_{j, k, l} \frac{g_{jkl} (x_{ji_1} u_{jl} - \bar{x}_{i_1} \bar{u}) (x_{ji_2} - \bar{x}_{i_2})}{\sigma_j^2} & 1 \leq i_1 \leq r, \quad r+2 \leq i_2 \leq r+s+1 \\ \sum_{j, k, l} \frac{g_{jkl} (x_{ji_1} u_{jl} - \bar{x}_{i_1} \bar{u}) (u_{jk} - \bar{u})}{\sigma_j^2} & r+2 \leq i_1 \leq r+s+1, \quad i_2 = r+1 \\ \sum_{j, k, l} \frac{g_{jkl} (x_{ji_1} u_{jl} - \bar{x}_{i_1} \bar{u}) (x_{ji_2} u_{jk} - \bar{x}_{i_2} \bar{u})}{\sigma_j^2} & r+2 \leq i_1 \leq r+s+1, \quad r+2 \leq i_2 \leq r+s+1 \end{cases} \quad (55)$$

当 σ_j^2 已知时, 可以证明 $\hat{a}_0, \hat{a}_1, \dots, \hat{a}_r$ 和 $\hat{b}_0, \hat{b}_1, \dots, \hat{b}_s$ 是 $a_0, a_1, \dots, a_r$ 和 $b_0, b_1, \dots, b_s$ 的最佳无偏估计; 当 σ_j^2 未知时, 可先取 $\sigma_j^2 = 1$ 代入 (47)~(55) 求出 $\hat{a}_0, \hat{a}_1, \dots, \hat{a}_r$ 和 $\hat{b}_0, \hat{b}_1, \dots, \hat{b}_s$ 的初值, 并计算 σ_j 的估计值 $\hat{\sigma}_j = \hat{b}_0 + \hat{b}_1 x_1 + \dots + \hat{b}_s x_s$, 然后反复迭代计算, 直到 $\hat{\sigma}_j$ 稳定为止。大量 Monte Carlo 模拟表明, 当数据总量稍大时, $\hat{a}_0, \hat{a}_1, \dots, \hat{a}_r$ 和 $\hat{b}_0, \hat{b}_1, \dots, \hat{b}_s$ 即可用正态分布近似, 它们的协方差矩阵为:

$$\text{Cov}(\hat{a}_0, \hat{a}_1, \dots, \hat{a}_r, \hat{b}_0, \hat{b}_1, \dots, \hat{b}_s)^T = C = Z^{-1} \quad (56)$$

式中矩阵 Z 的元素 $z_{i_1 i_2}$ 为:

$$z_{i_1 i_2} = \begin{cases} \sum_{j, k, l} \frac{g_{jkl} X_{j(i_1-1)} X_{j(i_2-1)}}{\sigma_j^2} & 1 \leq i_1 \leq r+1, \quad 1 \leq i_2 \leq r+1 \\ \sum_{j, k, l} \frac{g_{jkl} X_{j(i_1-1)} X_{j(i_2-2)} u_{jk}}{\sigma_j^2} & 1 \leq i_1 \leq r+1, \quad r+2 \leq i_2 \leq r+s+2 \\ \sum_{j, k, l} \frac{g_{jkl} X_{j(i_1-r-2)} X_{j(i_2-2)} u_{jk} u_{jl}}{\sigma_j^2} & r+2 \leq i_1 \leq r+s+2, \quad r+2 \leq i_2 \leq r+s+2 \end{cases} \quad (57)$$

如果不进行迭代计算, 而直接在式 (47)~(55) 中取 $\sigma_j^2 = 1$ 时, 则式 (45) 和 (46) 给出 $a_0, a_1, \dots, a_r$ 和 $b_0, b_1, \dots, b_s$ 的无偏估计, 且它们的协方差矩阵为:

$$\text{Cov}(\hat{a}_0, \hat{a}_1, \dots, \hat{a}_r, \hat{b}_0, \hat{b}_1, \dots, \hat{b}_s)^T = C = E^{-1} D E^{-1} \quad (58)$$

式中矩阵 E 和 D 的元素 $e_{i_1 i_2}$ 和 $d_{i_1 i_2}$ 分别为:

$$e_{i_1 i_2} = \begin{cases} \sum_{j, k, l} g_{jkl} X_{j(i_1-1)} X_{j(i_2-1)} & 1 \leq i_1 \leq r+1, \quad 1 \leq i_2 \leq r+1 \\ \sum_{j, k, l} g_{jkl} X_{j(i_1-1)} X_{j(i_2-r-2)} u_{jk} & 1 \leq i_1 \leq r+1, \quad r+2 \leq i_2 \leq r+s+2 \\ \sum_{j, k, l} g_{jkl} X_{j(i_1-r-2)} X_{j(i_2-r-2)} u_{jk} u_{jl} & r+2 \leq i_1 \leq r+s+2, \quad r+2 \leq i_2 \leq r+s+2 \end{cases} \quad (59)$$

$$d_{i_1 i_2} = \begin{cases} \sum_{j, k, l} \hat{\sigma}_j^2 g_{jkl} X_{j(i_1-1)} X_{j(i_2-1)} & 1 \leq i_1 \leq r+1, \quad 1 \leq i_2 \leq r+1 \\ \sum_{j, k, l} \hat{\sigma}_j^2 g_{jkl} X_{j(i_1-1)} X_{j(i_2-r-2)} u_{jk} & 1 \leq i_1 \leq r+1, \quad r+2 \leq i_2 \leq r+s+2 \\ \sum_{j, k, l} \hat{\sigma}_j^2 g_{jkl} X_{j(i_1-r-2)} X_{j(i_2-r-2)} u_{jk} u_{jl} & r+2 \leq i_1 \leq r+s+2, \quad r+2 \leq i_2 \leq r+s+2 \end{cases} \quad (60)$$

式中 $x_{j0} = 1, j = 1, 2, \dots, m_0$

3.2 多元回归方程和百分位值曲线的置信限估计

随机变量 y 的回归方程为:

$$\hat{y}(x) = \hat{a}_0 + \hat{a}_1 x_1 + \dots + \hat{a}_r x_r \quad (61)$$

回归曲线的置信度为 γ 的单侧置信下限和单侧置信上限分别为:

$$\hat{y}_L(x) = \hat{a}_0 + \hat{a}_1 X_1 + \cdots + \hat{a}_r X_r - u_{\gamma} \sqrt{C_{11} + 2 \sum_{i=1}^r C_{1(i+1)} X_i + \sum_{k,l=1}^r C_{(k+1)(l+1)} X_k X_l} \quad (62)$$

$$\hat{y}_U(x) = \hat{a}_0 + \hat{a}_1 X_1 + \cdots + \hat{a}_r X_r + u_{\gamma} \sqrt{C_{11} + 2 \sum_{i=1}^r C_{1(i+1)} X_i + \sum_{k,l=1}^r C_{(k+1)(l+1)} X_k X_l} \quad (63)$$

标准差 $\sigma(x)$ 的估计量为:

$$\hat{\sigma}(x) = \hat{b}_0 + \hat{b}_1 X_1 + \cdots + \hat{b}_s X_s \quad (64)$$

$\sigma(x)$ 的置信度为 γ 的单侧置信下限和单侧置信上限分别为:

$$\hat{\alpha}(x) = \hat{b}_0 + \hat{b}_1 X_1 + \cdots + \hat{b}_s X_s - u_{\gamma} \sqrt{C_{(r+2)(r+2)} + 2 \sum_{i=1}^s C_{(r+2)(i+r+2)} X_i + \sum_{k,l=1}^s C_{(k+r+2)(l+r+2)} X_k X_l} \quad (65)$$

$$\hat{\alpha}_U(x) = \hat{b}_0 + \hat{b}_1 X_1 + \cdots + \hat{b}_s X_s + u_{\gamma} \sqrt{C_{(r+2)(r+2)} + 2 \sum_{i=1}^s C_{(r+2)(i+r+2)} X_i + \sum_{k,l=1}^s C_{(k+r+2)(l+r+2)} X_k X_l} \quad (66)$$

随机变量 y 的概率为 P 的百分位值 $y_P(x)$ 的无偏估计量由下式给出:

$$y_P(x) = \hat{a}_0 + \hat{a}_1 X_1 + \cdots + \hat{a}_r X_r + u_P(\hat{b}_0 + \hat{b}_1 X_1 + \cdots + \hat{b}_s X_s) \quad (67)$$

$y_P(x)$ 的置信度为 γ 的单侧置信下限 $y_{PL}(x)$ 和单侧置信上限 $y_{PU}(x)$ 分别为:

$$y_{PL}(x) = \hat{a}_0 + \hat{a}_1 X_1 + \cdots + \hat{a}_r X_r + u_P(\hat{b}_0 + \hat{b}_1 X_1 + \cdots + \hat{b}_s X_s) - u_{\gamma} \left\{ C_{11} + 2u_P C_{1(r+2)} + u_P^2 C_{(r+2)(r+2)} + 2 \sum_{i=1}^r (C_{1(i+1)} + u_P C_{(i+1)(r+2)}) X_i + 2u_P \sum_{i=1}^s (C_{1(i+r+2)} + u_P C_{(r+2)(i+r+2)}) X_i + \sum_{k,l=1}^r C_{(k+1)(l+1)} X_k X_l + u_P^2 \sum_{k,l=1}^s C_{(k+r+2)(l+r+2)} X_k X_l + 2u_P \sum_{k=1}^r \sum_{l=1}^s C_{(k+1)(l+r+2)} X_k X_l \right\}^{1/2} \quad (68)$$

$$y_{PU}(x) = \hat{a}_0 + \hat{a}_1 X_1 + \cdots + \hat{a}_r X_r + u_P(\hat{b}_0 + \hat{b}_1 X_1 + \cdots + \hat{b}_s X_s) + u_{\gamma} \left\{ C_{11} + 2u_P C_{1(r+2)} + u_P^2 C_{(r+2)(r+2)} + 2 \sum_{i=1}^r (C_{1(i+1)} + u_P C_{(i+1)(r+2)}) X_i + 2u_P \sum_{i=1}^s (C_{1(i+r+2)} + u_P C_{(r+2)(i+r+2)}) X_i + \sum_{k,l=1}^r C_{(k+1)(l+1)} X_k X_l + u_P^2 \sum_{k,l=1}^s C_{(k+r+2)(l+r+2)} X_k X_l + 2u_P \sum_{k=1}^r \sum_{l=1}^s C_{(k+1)(l+r+2)} X_k X_l \right\}^{1/2} \quad (69)$$

如果将式(45)~式(60)中的 u_{jk} 和 v_{jkl} 改为标准极值分布顺序统计量的均值和协方差, 则它们对极值分布仍然成立, 并且可以类似给出极值分布和威布尔分布多元异方差回归方程和百分位值曲线估计。

4 结 论

(1) 提出截尾数据异方差回归分析方法, 将只适用于同方差和完全数据的传统回归分析推广到异方差和截尾数据情况。

(2) 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。

(3) 建立回归曲线和百分位值的置信限估计。

(4) 本文方法不仅适用于正态分布, 而且也适用于威布尔分布和极值分布。

(5) 建立多元异方差回归分析方法。

(6) 本文方法能够将不同状态下的试验数据作为一个整体进行统计分析, 其估计精度明显优于只能在各自状态下进行统计分析的成组试验法, 可以节省大量试样。

参考文献

- [1] Olive Jean Dunn, Virginia A. Clark. Applied Statistics: A Analysis of Variance 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, 25(4): 465- 469.
- [4] 黄伟, 傅惠民. 威布尔分布多元回归分析方法[J]. 航空动力学报, 2003, 25(5): 583- 586.