

Heteroscedastic regression analysis method for mixed data

FU Hui-min, YUE Xiao-rui

(Research Center of Small Sample Technology,
Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Abstract: The heteroscedastic regression model was established and the heteroscedastic regression analysis method was presented for mixed data composed of complete data, type-I censored data and type-II censored data from the location-scale distribution. The best unbiased estimations of regression coefficients, as well as the confidence limits of the location parameter and scale parameter were given. Furthermore, the point estimations and confidence limits of percentiles were obtained. Thus, the traditional multiple regression analysis method which is only suitable to the complete data from normal distribution can be extended to the cases of heteroscedastic mixed data and the location-scale distribution. So the presented method has a broad range of promising applications.

Key words: heteroscedastic regression analysis; censored data; performance test; life prediction; reliability analysis

CLC number: TB301; O211

Document code: A

Introduction

In multiple regression analysis, the model is referred to as heteroscedastic regression model if the variance of disturbance varies with the observations. Heteroscedasticity is a very common phenomenon in engineering technology and econometrics. For example, in the test of $S-N$ and $P-S-N$ curves for materials, the dispersity of life augments as the stress level decreases, which is indicated by a great number of experimental results^[1]. This means that the disturbances of life at different stresses have unequal variances. Heteroscedasticity of data can be tested by several test methods^[2-3], e. g. graph method, Spearman rank correlation coefficients test, Glejser test, Goldfeld-Quandt test. It is of great significance to develop a statistical inference for the regression models with heteroscedasticity.

In the previous works, FU and HUANG^[4] established the best unbiased integral estimation

of the location-scale distribution parameters for type-II censored data with heteroscedasticity. FU and YUE^[5-6] presented the regression analysis methods for mixed data which are composed of complete data, type-I censored data and type-II censored data with homoscedasticity. In this study, the heteroscedastic regression analysis method for mixed data is put forward under the location-scale distribution. Thus, the traditional multiple regression analysis is extended to the cases of heteroscedastic mixed data and location-scale distribution.

1 Heteroscedastic regression analysis for mixed data

Location-scale distribution is widely used in engineering. Many popular life distributions belong to the location-scale distribution family, including Weibull distribution, the extreme value distribution, normal distribution, and log-normal distribution. The random variable y is sup-

Received: 2010-12-26; **revision received:** 2011-03-17

Foundation item: Supported by the National Natural Science Foundation of China (Grant No. 10472006)

E-mail: fuhuimin@263.net

posed to follow the location-scale distribution with the cumulative distribution function (cdf)

$$F\left[\frac{y-\mu(\mathbf{a}, \mathbf{x})}{\sigma(\mathbf{b}, \mathbf{x})}\right], \quad -\infty < y < +\infty \quad (1)$$

where $\mu(\mathbf{a}, \mathbf{x})$ and $\sigma(\mathbf{b}, \mathbf{x})$ are the location parameter and scale parameter at the state (independent variable) $\mathbf{x} = (x_1, x_2, \dots, x_r)^T$, respectively. The column vectors $\mathbf{a} = (a_0, a_1, \dots, a_h)^T$ and $\mathbf{b} = (b_0, b_1, \dots, b_s)^T$ are distribution parameters to be estimated. Let

$$y^\circ = \frac{y - \mu(\mathbf{a}, \mathbf{x})}{\sigma(\mathbf{b}, \mathbf{x})} \quad (2)$$

then y° is the random variable from the standard location-scale distribution with the cdf $F(y^\circ)$ and the probability density function (pdf) $f(y^\circ)$, where $F(y^\circ)$ and $f(y^\circ)$ do not rely on any unknown parameters.

In many situations, $\mu(\mathbf{a}, \mathbf{x})$ and $\sigma(\mathbf{b}, \mathbf{x})$ can be expressed as the functions of $\mathbf{x}$ in the following forms

$$\mu(\mathbf{a}, \mathbf{x}) = \sum_{j=0}^h a_j \varphi_j(\mathbf{x}) \quad (3)$$

$$\sigma(\mathbf{b}, \mathbf{x}) = \sum_{j=0}^s b_j \psi_j(\mathbf{x}) \quad (4)$$

where $\varphi_j(\mathbf{x})$ ($j = 0, 1, \dots, h$) and $\psi_j(\mathbf{x})$ ($j = 0, 1, \dots, s$) are known functions of the state $\mathbf{x}$. The relationship between $\mu(\mathbf{a}, \mathbf{x})$, $\sigma(\mathbf{b}, \mathbf{x})$ and $\mathbf{a}, \mathbf{b}, \mathbf{x}$ can be derived from many approaches, such as the physical model of the investigated issues, the results of relative theoretical derivation, or the empirical formulas. Moreover, the relationship in Eqs. (3) and (4) can be tested by referring to Ref. [7].

1.1 Heteroscedastic regression model

Suppose that the test was performed at m states. A random sample of n_i specimens were put on test at the i th state $\mathbf{x}_i = (x_{1i}, x_{2i}, \dots, x_{ri})^T$, and q_i specimens were observed to fail. The ordered failure data were denoted by

$y_{i1} \leq y_{i2} \leq \dots \leq y_{iq_i}$. The mixed data obtained at m states were composed of three parts:

1) Complete data obtained at the first m_1 states $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{m_1}$. At these states, $q_i = n_i$, $i = 1, 2, \dots, m_1$.

2) Type-II censored data obtained at the next m_2 states $\mathbf{x}_{m_1+1}, \mathbf{x}_{m_1+2}, \dots, \mathbf{x}_{m_1+m_2}$. At these states, the largest failure data equals to the censoring time, and $q_i < n_i$, $i = m_1 + 1, m_1 + 2, \dots, m_1 + m_2$.

3) Type-I censored data obtained at the last m_3 states $\mathbf{x}_{m_1+m_2+1}, \mathbf{x}_{m_1+m_2+2}, \dots, \mathbf{x}_{m_1+m_2+m_3}$, and $m_1 + m_2 + m_3 = m$. At these states, the censoring time denoted by $y_{i(q_i+1)}$ is larger than the largest failure time y_{iq_i} .

According to Ref. [8], $y_{i(q_i+1)}$ ($i = m_1 + m_2 + 1, m_1 + m_2 + 2, \dots, m$) can be regarded as an observed value of the q_i th interval statistic from the location-scale distribution with a sample of size n_i .

According to Eq. (2), we can obtain

$$y_{ik} = \mu(\mathbf{a}, \mathbf{x}_i) + \sigma(\mathbf{b}, \mathbf{x}_i) y_{ik}^\circ$$

$$\text{when } i = 1, 2, \dots, m_1 + m_2,$$

$$k = 1, 2, \dots, q_i;$$

$$\text{when } i = m_1 + m_2 + 1, m_1 + m_2 + 2, \dots, m,$$

$$k = 1, 2, \dots, q_i + 1 \quad (5)$$

where y_{ik}° ($i = 1, 2, \dots, m$; $k = 1, 2, \dots, q_i$) is the k th order statistic under the standard location-scale distribution with a sample of size n_i , and $y_{i(q_i+1)}^\circ$ ($i = m_1 + m_2 + 1, m_1 + m_2 + 2, \dots, m$) is the q_i th interval statistic from the same sample. Let u_{ik} and v_{ikl} denote the expectation and covariance of order statistics from standard location-scale distribution, respectively, i. e.

$$E(y_{ik}^\circ) = u_{ik} \quad (6)$$

$$\text{Cov}(y_{ik}^\circ, y_{il}^\circ) = v_{ikl} \quad (7)$$

in which u_{ik} and v_{ikl} can be calculated by the following formulas:

$$u_{ik} = \frac{n_i!}{(k-1)!(n_i-k)!} \int_{-\infty}^{\infty} y^\circ f(y^\circ) [F(y^\circ)]^{k-1} [1-F(y^\circ)]^{n_i-k} dy^\circ$$

$$i = 1, 2, \dots, m; \quad k = 1, 2, \dots, q_i \quad (8)$$

$$u_{i(q_i+1)} = \frac{(n_i+1)!}{q_i!(n_i-q_i)!} \int_{-\infty}^{\infty} y^\circ f(y^\circ) [F(y^\circ)]^{q_i} [1-F(y^\circ)]^{n_i-q_i} dy^\circ$$

$$i = m_1 + m_2 + 1, m_1 + m_2 + 2, \dots, m \quad (9)$$

$$\nu_{ikl} = E(y_{ik}^{\circ}, y_{il}^{\circ}) - u_{ik} u_{il} \quad (10)$$

$$E(y_{ik}^{\circ}, y_{il}^{\circ}) = \frac{n_i!}{(k-1)!(l-k-1)!(n_i-l)!} \cdot \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y_{ik}^{\circ} y_{il}^{\circ} f(y_{ik}^{\circ}) f(y_{il}^{\circ}) [F(y_{ik}^{\circ})]^{k-1} [F(y_{il}^{\circ}) - F(y_{ik}^{\circ})]^{l-k-1} [1 - F(y_{il}^{\circ})]^{n_i-l} dy_{ik}^{\circ} dy_{il}^{\circ} \quad (11)$$

$$i = 1, 2, \dots, m; \quad k, l = 1, 2, \dots, q_i$$

$$E(y_{ik}^{\circ}, y_{i(q_i+1)}^{\circ}) = \frac{(n_i+1)!}{(k-1)!(q_i-k)!(n_i-q_i)!} \cdot \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y_{ik}^{\circ} y_{i(q_i+1)}^{\circ} f(y_{ik}^{\circ}) f(y_{i(q_i+1)}^{\circ}) [F(y_{ik}^{\circ})]^{k-1} [F(y_{i(q_i+1)}^{\circ}) - F(y_{ik}^{\circ})]^{q_i-k} [1 - F(y_{i(q_i+1)}^{\circ})]^{n_i-q_i} dy_{ik}^{\circ} dy_{i(q_i+1)}^{\circ} \quad (12)$$

$$i = m_1 + m_2 + 1, m_1 + m_2 + 2, \dots, m; \quad k = 1, 2, \dots, q_i + 1$$

Defining the disturbance corresponding to $\mathbf{x}_i$ as

$$\epsilon_{ik} = \sigma(\mathbf{b}, \mathbf{x}_i)(y_{ik}^{\circ} - u_{ik}) \quad (13)$$

then from Eqs. (6) and (7) we can know that

$$E(\epsilon_{ik}) = 0 \quad (14)$$

$$\text{Cov}(\epsilon_{ik}, \epsilon_{il}) = \sigma^2(\mathbf{b}, \mathbf{x}_i) \nu_{ikl} \quad (15)$$

Substituting Eqs. (3), (4) and (13) into Eq. (5), the heteroscedastic regression model for mixed data can be given by

$$y_{ik} = \sum_{j=0}^h a_j \varphi_j(\mathbf{x}_i) + u_{ik} \sum_{j=0}^s b_j \psi_j(\mathbf{x}_i) + \epsilon_{ik}$$

when $i = 1, 2, \dots, m_1 + m_2$,

$$k = 1, 2, \dots, q_i;$$

when $i = m_1 + m_2 + 1, m_1 + m_2 + 2, \dots, m$,

$$k = 1, 2, \dots, q_i + 1 \quad (16)$$

$$E(y_{ik}) = \sum_{j=0}^h a_j \varphi_j(\mathbf{x}_i) + u_{ik} \sum_{j=0}^s b_j \psi_j(\mathbf{x}_i) \quad (17)$$

$$\text{Cov}(y_{ik}, y_{il}) = \sigma^2(\mathbf{b}, \mathbf{x}_i) \nu_{ikl} \quad (18)$$

For simplicity, let

$$\boldsymbol{\theta} = (a_0, a_1, \dots, a_h, b_0, b_1, \dots, b_s)^T \quad (19)$$

$$\mathbf{y}_i = \begin{cases} (y_{i1}, y_{i2}, \dots, y_{iq_i})^T & i = 1, 2, \dots, m_1 + m_2 \\ (y_{i1}, y_{i2}, \dots, y_{i(q_i+1)})^T & i = m_1 + m_2 + 1, m_1 + m_2 + 2, \dots, m \end{cases} \quad (20)$$

$$\boldsymbol{\epsilon}_i = \begin{cases} (\epsilon_{i1}, \epsilon_{i2}, \dots, \epsilon_{iq_i})^T & i = 1, 2, \dots, m_1 + m_2 \\ (\epsilon_{i1}, \epsilon_{i2}, \dots, \epsilon_{i(q_i+1)})^T & i = m_1 + m_2 + 1, m_1 + m_2 + 2, \dots, m \end{cases} \quad (21)$$

$$\mathbf{V}_i = \begin{cases} [\nu_{ikl}]_{q_i \times q_i} & i = 1, 2, \dots, m_1 + m_2 \\ [\nu_{ikl}]_{(q_i+1) \times (q_i+1)} & i = m_1 + m_2 + 1, m_1 + m_2 + 2, \dots, m \end{cases} \quad (22)$$

$$\mathbf{Z}_i = \begin{bmatrix} \varphi_0(\mathbf{x}_i) & \cdots & \varphi_h(\mathbf{x}_i) & \psi_0(\mathbf{x}_i) u_{i1} & \cdots & \psi_s(\mathbf{x}_i) u_{i1} \\ \varphi_0(\mathbf{x}_i) & \cdots & \varphi_h(\mathbf{x}_i) & \psi_0(\mathbf{x}_i) u_{i2} & \cdots & \psi_s(\mathbf{x}_i) u_{i2} \\ \vdots & & \vdots & \vdots & & \vdots \\ \varphi_0(\mathbf{x}_i) & \cdots & \varphi_h(\mathbf{x}_i) & \psi_0(\mathbf{x}_i) u_{iq_i} & \cdots & \psi_s(\mathbf{x}_i) u_{iq_i} \end{bmatrix} \quad i = 1, 2, \dots, m_1 + m_2 \quad (23)$$

$$\mathbf{Z}_i = \begin{bmatrix} \varphi_0(\mathbf{x}_i) & \cdots & \varphi_h(\mathbf{x}_i) & \psi_0(\mathbf{x}_i) u_{i1} & \cdots & \psi_s(\mathbf{x}_i) u_{i1} \\ \varphi_0(\mathbf{x}_i) & \cdots & \varphi_h(\mathbf{x}_i) & \psi_0(\mathbf{x}_i) u_{i2} & \cdots & \psi_s(\mathbf{x}_i) u_{i2} \\ \vdots & & \vdots & \vdots & & \vdots \\ \varphi_0(\mathbf{x}_i) & \cdots & \varphi_h(\mathbf{x}_i) & \psi_0(\mathbf{x}_i) u_{i(q_i+1)} & \cdots & \psi_s(\mathbf{x}_i) u_{i(q_i+1)} \end{bmatrix} \quad i = m_1 + m_2 + 1, m_1 + m_2 + 2, \dots, m \quad (24)$$

then the heteroscedastic regression model

$$\mathbf{y}_i = \mathbf{Z}_i \boldsymbol{\theta} + \boldsymbol{\epsilon}_i \quad (25)$$

Eqs. (16) ~ (18) can be rewritten in matrix

$$E(\mathbf{y}_i) = \mathbf{Z}_i \boldsymbol{\theta} \quad (26)$$

forms as

$$\text{Cov}(\mathbf{y}_i) = \sigma^2(\mathbf{b}, \mathbf{x}_i) \mathbf{V}_i \quad (27)$$

1.2 Unbiased estimations of parameters

Using the Gauss-Markov theorem, let the residual sum of squares is

$$Q = \sum_{i=1}^m \frac{(\mathbf{y}_i - \mathbf{Z}_i \boldsymbol{\theta})^T \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{Z}_i \boldsymbol{\theta})}{\sigma^2(\mathbf{b}, \mathbf{x}_i)} \quad (28)$$

Letting

$$\frac{\partial Q}{\partial \boldsymbol{\theta}} = 0 \quad (29)$$

then the best linear unbiased estimation of $\boldsymbol{\theta}$ can be given by

$$\hat{\boldsymbol{\theta}} = (\hat{a}_0, \hat{a}_1, \dots, \hat{a}_h, \hat{b}_0, \hat{b}_1, \dots, \hat{b}_s)^T = \left[\sum_{i=1}^m \frac{\mathbf{Z}_i^T \mathbf{V}_i^{-1} \mathbf{Z}_i}{\sigma^2(\mathbf{b}, \mathbf{x}_i)} \right]^{-1} \left[\sum_{i=1}^m \frac{\mathbf{Z}_i^T \mathbf{V}_i^{-1} \mathbf{y}_i}{\sigma^2(\mathbf{b}, \mathbf{x}_i)} \right] \quad (30)$$

And the covariance matrix of $\hat{\boldsymbol{\theta}}$ is

$$\text{Cov}(\hat{\boldsymbol{\theta}}) = \left[\sum_{i=1}^m \frac{\mathbf{Z}_i^T \mathbf{V}_i^{-1} \mathbf{Z}_i}{\sigma^2(\mathbf{b}, \mathbf{x}_i)} \right]^{-1} \quad (31)$$

If $\sigma^2(\mathbf{b}, \mathbf{x}_i)$ ($i=1, 2, \dots, m$) is unknown in advance, the unbiased estimation of $\boldsymbol{\theta}$ should be calculated by the following steps:

1) Let the initial value of $\sigma^2(\mathbf{b}, \mathbf{x}_i)$ ($i=1, 2, \dots, m$) in Eq. (30) equals 1, then the initial estimated value of $\boldsymbol{\theta} = (a_0, a_1, \dots, a_h, b_0, b_1, \dots, b_s)^T$ can be obtained.

2) Substituting the initial estimated value of $b_0, b_1, \dots, b_s$ into Eq. (4), the iterative estimated value of $\sigma^2(\mathbf{b}, \mathbf{x}_i)$ can be obtained.

3) Substituting the obtained iterative estimated value of $\sigma^2(\mathbf{b}, \mathbf{x}_i)$ into Eq. (30), and the process is iteratively repeated until the estimated value of $\sigma^2(\mathbf{b}, \mathbf{x}_i)$ become stable.

It has been testified by enormous process of computing that the specified iterative precision can be reached after a few iterations.

1.3 Confidence limits of location parameter and scale parameter

According to Eqs. (3) and (4), the unbiased point estimation of $\mu(\mathbf{a}, \mathbf{x})$ and $\sigma(\mathbf{b}, \mathbf{x})$ are

$$\hat{\mu}(\mathbf{a}, \mathbf{x}) = \sum_{j=0}^h \hat{a}_j \varphi_j(\mathbf{x}) \quad (32)$$

$$\hat{\sigma}(\mathbf{b}, \mathbf{x}) = \sum_{j=0}^s \hat{b}_j \psi_j(\mathbf{x}) \quad (33)$$

Large amounts of Monte-Carlo simulations

had been conducted and the results showed that $\hat{\boldsymbol{\theta}}$ approximately submits the normal distribution if the total number of failures $\sum_{i=1}^m q_i$ is great in quantity. Therefore, the one-side upper confidence limits and lower confidence limits of $\mu(\mathbf{a}, \mathbf{x})$ and $\sigma(\mathbf{b}, \mathbf{x})$ can be given by

$$\mu_{\text{up}}(\mathbf{x}) = \sum_{j=0}^h \hat{a}_j \varphi_j(\mathbf{x}) + u_{\gamma} \sqrt{\sum_{j_1, j_2=1}^{h+1} c_{j_1 j_2} \varphi_{j_1-1}(\mathbf{x}) \varphi_{j_2-1}(\mathbf{x})} \quad (34)$$

$$\mu_{\text{low}}(\mathbf{x}) = \sum_{j=0}^h \hat{a}_j \varphi_j(\mathbf{x}) - u_{\gamma} \sqrt{\sum_{j_1, j_2=1}^{h+1} c_{j_1 j_2} \varphi_{j_1-1}(\mathbf{x}) \varphi_{j_2-1}(\mathbf{x})} \quad (35)$$

$$\sigma_{\text{up}}(\mathbf{x}) = \sum_{j=0}^s \hat{b}_j \psi_j(\mathbf{x}) + u_{\gamma} \sqrt{\sum_{j_1, j_2=h+2}^{h+s+2} c_{j_1 j_2} \psi_{j_1-h-2}(\mathbf{x}) \psi_{j_2-h-2}(\mathbf{x})} \quad (36)$$

$$\sigma_{\text{low}}(\mathbf{x}) = \sum_{j=0}^s \hat{b}_j \psi_j(\mathbf{x}) - u_{\gamma} \sqrt{\sum_{j_1, j_2=h+2}^{h+s+2} c_{j_1 j_2} \psi_{j_1-h-2}(\mathbf{x}) \psi_{j_2-h-2}(\mathbf{x})} \quad (37)$$

where γ is the degree of confidence, u_{γ} is the 100 γ th percentile of the standard normal distribution, and

$$[c_{j_1 j_2}]_{(h+s+2) \times (h+s+2)} = \left[\sum_{i=1}^m \frac{\mathbf{Z}_i^T \mathbf{V}_i^{-1} \mathbf{Z}_i}{\sigma^2(\mathbf{b}, \mathbf{x}_i)} \right]^{-1} \quad (38)$$

1.4 Point estimation and confidence limits of percentiles

The unbiased estimation of the 100Pth percentile of the random variable y at the state $\mathbf{x}$ is

$$\hat{y}_P(\mathbf{x}) = \sum_{j=0}^h \hat{a}_j \varphi_j(\mathbf{x}) + F^{-1}(P) \sum_{j=0}^s \hat{b}_j \psi_j(\mathbf{x}) \quad (39)$$

Its one-side upper confidence limit $y_{\text{Pup}}(\mathbf{x})$ and lower confidence limit $y_{\text{Plow}}(\mathbf{x})$ are

$$y_{\text{Pup}}(\mathbf{x}) = \sum_{j=0}^h \hat{a}_j \varphi_j(\mathbf{x}) + F^{-1}(P) \sum_{j=0}^s \hat{b}_j \psi_j(\mathbf{x}) + u_{\gamma} \sqrt{\omega} \quad (40)$$

$$y_{\text{Plow}}(\mathbf{x}) = \sum_{j=0}^h \hat{a}_j \varphi_j(\mathbf{x}) + F^{-1}(P) \sum_{j=0}^s \hat{b}_j \psi_j(\mathbf{x}) - u_\gamma \sqrt{\omega} \quad (41)$$

where

$$\begin{aligned} \omega &= \sum_{j_1, j_2=1}^{h+1} c_{j_1 j_2} \varphi_{j_1-1}(\mathbf{x}) \varphi_{j_2-1}(\mathbf{x}) + \\ &F^{-1}(P) \sum_{j_1=1}^{h+1} \sum_{j_2=h+2}^{h+s+2} c_{j_1 j_2} \varphi_{j_1-1}(\mathbf{x}) \psi_{j_2-h-2}(\mathbf{x}) + \\ &F^{-1}(P) \sum_{j_1=h+2}^{h+s+2} \sum_{j_2=1}^{h+1} c_{j_1 j_2} \varphi_{j_1-h-2}(\mathbf{x}) \psi_{j_2-1}(\mathbf{x}) + \\ &\left[F^{-1}(P) \right]^2 \sum_{j_1, j_2=h+2}^{h+s+2} c_{j_1 j_2} \psi_{j_1-h-2}(\mathbf{x}) \psi_{j_2-h-2}(\mathbf{x}) \end{aligned} \quad (42)$$

2 Heteroscedastic regression analysis method for several widely used distributions

2.1 Normal distribution case

Suppose the random variable y follows the normal distribution with the cdf

$$F(y) = \Phi \left[\frac{y - \mu(\mathbf{a}, \mathbf{x})}{\sigma(\mathbf{b}, \mathbf{x})} \right] \quad (43)$$

where $\mu(\mathbf{a}, \mathbf{x})$ and $\sigma(\mathbf{b}, \mathbf{x})$ are, respectively, the mean value and the standard deviation at the state $\mathbf{x} = (x_1, x_2, \dots, x_r)^T$, and $\Phi(\cdot)$ denotes the cdf of the standard normal distribution. If the relationship between $\mu(\mathbf{a}, \mathbf{x})$, $\sigma(\mathbf{b}, \mathbf{x})$ and the state $\mathbf{x}$ can be expressed as the forms of Eqs. (3) and (4), the best unbiased estimation of $a_0, \dots, a_h$ and $b_0, \dots, b_s$ can be obtained by Eq. (30). The u_{ik} and v_{ikl} in Eqs. (22)~(24) are the expectations and covariances of the order statistics from standard normal distribution, respectively. The expectations u_{ik} and covariances v_{ikl} of order statistics for sample size up to 20 were tabulated by Sarhan and Greenberg^[9]. Then the tables were extended to sample size 50 by Tietjen et al^[10]. Later more precise values of these tables were given by Parrish^[11-12]. Recently, FU and LIN^[13] studied the computation of u_{ik} and v_{ikl} for large sample size. Consequently, u_{ik} and v_{ikl} can be obtained by consulting the tables in Refs. [11-12] when the sample size is smaller than 50,

and can be calculated by referring to Ref. [13] when the sample size is larger than 50.

The confidence limits of the mean value $\mu(\mathbf{a}, \mathbf{x})$ and the standard deviation $\sigma(\mathbf{b}, \mathbf{x})$ can be obtained according to section 1.3. The point estimation and confidence limits of percentiles can be computed in terms of section 1.4 with $F^{-1}(P)$ in Eqs. (39)~(42) substituted by $\Phi^{-1}(P)$.

2.2 Extreme value distribution case

Suppose the random variable y follows the extreme value distribution with the cdf

$$F(y) = 1 - \exp \left\{ -\exp \left[\frac{y - \mu(\mathbf{a}, \mathbf{x})}{\sigma(\mathbf{b}, \mathbf{x})} \right] \right\} \quad (44)$$

where $\mu(\mathbf{a}, \mathbf{x})$ and $\sigma(\mathbf{b}, \mathbf{x})$ are, respectively, the location parameter and the scale parameter at the state $\mathbf{x} = (x_1, x_2, \dots, x_r)^T$. If the relationship between $\mu(\mathbf{a}, \mathbf{x})$, $\sigma(\mathbf{b}, \mathbf{x})$ and the state $\mathbf{x}$ can be expressed as the forms of Eqs. (3) and (4), the best unbiased estimation of $a_0, \dots, a_h$ and $b_0, \dots, b_s$ can be obtained by Eq. (30). The u_{ik} and v_{ikl} in Eqs. (22)~(24) are the expectations and covariances of the order statistics from standard extreme value distribution, respectively. Balakrishnan and Chan^[14] tabulated the expectations and covariances of order statistics for sample size up to 30. Therefore, u_{ik} and v_{ikl} can be obtained by consulting the tables in Ref. [14] when the sample size is smaller than 30, and can be calculated by referring to Ref. [13] when the sample size is large.

The confidence limits of the location parameter $\mu(\mathbf{a}, \mathbf{x})$ and the scale parameter $\sigma(\mathbf{b}, \mathbf{x})$ can be obtained according to section 1.3. The point estimation and confidence limits of percentiles can be computed according to section 1.4, with $F^{-1}(P)$ in Eqs. (39)~(42) substituted by $\ln \ln \frac{1}{1-P}$.

2.3 Weibull distribution case

Suppose the random variable t follows Weibull distribution with the cdf

$$F(t) = 1 - \exp \left\{ -\left[\frac{t}{\beta(\mathbf{x})} \right]^{\alpha(\mathbf{x})} \right\} \quad (45)$$

where $\alpha(\mathbf{x})$ and $\beta(\mathbf{x})$ are the shape parameter and characteristic life at the state $\mathbf{x} = (x_1, x_2, \dots,$

$x_r)^T$, respectively. Let $y = \ln t$, then y submits the extreme value distribution with the cdf given by Eq. (44), and

$$\begin{cases} \mu(\mathbf{a}, \mathbf{x}) = \ln \beta(\mathbf{x}) \\ \sigma(\mathbf{b}, \mathbf{x}) = 1/\alpha(\mathbf{x}) \end{cases} \quad (46)$$

If the $\mu(\mathbf{a}, \mathbf{x})$ and $\sigma(\mathbf{b}, \mathbf{x})$ in Eq. (46) can be expressed by Eqs. (3) and (4), the heteroscedastic regression analysis for mixed data from Weibull distribution will be converted into the heteroscedastic regression of the extreme value distribution case. Therefore, after logarithm transformation, the mixed data can be processed referring to section 2.2.

3 Conclusions

The main conclusions of this study are:

1) For mixed data composed of complete data and censored data, the heteroscedastic regression analysis method was presented under location-scale distribution. The unbiased estimations of regression coefficients, as well as the confidence limits of location parameter and scale parameter were obtained.

2) The point estimation and confidence limits of percentiles were given in the paper. Based on these, the confidence limits of percentages can be derived by using the equal principle of the confidence limit curve^[15].

3) The heteroscedastic regression analysis for mixed data from normal distribution, extreme value distribution and Weibull distribution were discussed in detail.

4) In the presented method, all the data including the complete data and different type of censored data were analyzed together. Numerous Monte-Carlo simulations showed that the present method for heteroscedastic data can reach a high precision.

5) The present method has a wide prospect of application, as it is applicable not only for the complete data from normal distribution but also for heteroscedastic mixed data from some other location-scale distributions.

References:

- [1] FU Huimin, LIU Chengrui. Small sample test method for S-N and P-S-N curves [J]. Journal of Mechanical Strength, 2006, 28(4): 552-555. (in Chinese)
- [2] BAI Xuemei. Testing methods of heteroscedasticity and comment [J]. Journal of Dongbei University of Finance and Economics, 2002, 24(6): 26-29. (in Chinese)
- [3] GONG Xiufang. The handling of heteroscedastic data in regression model [D]. Shanghai: East China Normal University, 2002. (in Chinese)
- [4] FU Huimin, HUANG Wei. Best unbiased integral estimate method [J]. Journal of Mechanical Strength, 2003, 25(3): 319-324. (in Chinese)
- [5] FU Huimin, YUE Xiaorui. Regression analysis method for mixed data [J]. Journal of Mechanical Strength, 2010, 32(6): 922-927. (in Chinese)
- [6] FU Huimin, YUE Xiaorui. Multiple regression analysis method for mixed data [J]. Journal of Aerospace Power, 2011, 26(1): 173-177. (in Chinese)
- [7] FU Huimin. Integral inference with maximum likelihood method [J]. Journal of Mechanical Strength, 2002, 24(1): 1-5. (in Chinese)
- [8] FU Huimin. Interval statistic and its distribution [J]. Journal of Mechanical Strength, 2005, 27(6): 752-757. (in Chinese)
- [9] Sarhan A E, Greenberg B G. Contributions to order statistics [M]. New York: John Wiley & Sons, 1962.
- [10] Tietjen G L, Kahaner D K, Beckman R J. Variances and covariances of the normal order statistics for sample sizes 2 to 50 [R]. Los Alamos, NM: Los Alamos Nat. Lab., LA-6482-MS, 1976.
- [11] Parrish R S. Computing expected values of normal order statistics [J]. Communications in Statistics-Simulation and Computation, 1992, 21(1): 57-70.
- [12] Parrish R S. Computing variances and covariances of normal order statistics [J]. Communications in Statistics-Simulation and Computation, 1992, 21(1): 71-101.
- [13] FU Huimin, LIN Fengchun. Computation and verification of means, variances and covariances for large sample order statistics [J]. Journal of Mechanical Strength, 2007, 29(1): 48-52. (in Chinese)
- [14] Balakrishnan N, Chan P S. Order statistics from extreme value distribution I: table of means, variances and covariances [J]. Communications in Statistics-Simulation and Computation, 1992, 21(4): 1199-1217.
- [15] FU Huimin. Interval estimation and hypothesis test for percentiles and percentages of continuous distributions [J]. Acta Aeronautica et Astronautica Sinica, 1993, 14(11): 569-577. (in Chinese)