应用概率统计 第三十卷
第四期 2014年8月

Chinese Journal of Applied Probability
and Statistics Vol.30 No.4 Aug. 2014

Estimation of Pareto Distribution from Interval Censored Observations *

Ding Bangjun

(*Department of Law and Politics, Kashi Teachers College, Kashi, 844006*)

(*School of Finance and Statistics, East China Normal University, Shanghai, 200241*)

Abstract

The problem of estimating the scale parameter in the Pareto distribution from interval censored observations is considered. Four kinds of estimators, including the maximum likelihood estimator and least square estimator, are evaluated. The variance of them are compared, and the numerical simulation results is also given.

Keywords: Pareto distribution, interval censored observations, parameter estimation.

AMS Subject Classification: 62N02.

§1. Introduction

Suppose that a random variable X follows Pareto distribution, i.e., the probability density function (PDF) of X is

$$f(x;\theta)=\begin{cases}\theta x^{-(\theta+1)} & \text{for } x\geq 1;\\ 0 & \text{for } x<1,\end{cases} \tag{1.1}$$

the parameter θ being positive. The corresponding cumulative distribution function (CDF) is

$$F(x;\theta)=\begin{cases}1-x^{-\theta} & \text{for } x\geq 1;\\ 0 & \text{for } x<1.\end{cases} \tag{1.2}$$

For example, if R denotes family income and R_0 denotes a base income, then over the range $R\geq R_0$, observed distributions of $X=R/R_0$ appears to be reasonably well described by (1.1). Now suppose that a random sample of size n is drawn from this probability distribution, and the observations are reported in the form of a set of interval censored

*The project was supported by the National Natural Science Foundation of China (44106460).

Received February 25, 2014. Revised July 7, 2014.

doi: 10.3969/j.issn.1001-4268.2014.04.007

data, such as (a_i, b_i), $i = 1, 2, \ldots, n$. Such data arise quite naturally in medical follow-up studies or in industrial life-testing. Turnbull (1976) and Chang and Yang (1987) dealt with the problem of estimating the underlying survival distribution, Ding (2012, 2008) gave a maximum likelihood estimation for some special laws under any interval censored observations. In this paper we consider a more restricted model than was assumed, and deal with the problem of estimating a parameter by maximum likelihood or least square.

The rest of the paper is organized as follows. In Section 2, we give a careful description of the interval censored model and discuss the maximum likelihood estimator (MLE) of the parameter θ. In Section 3, a non-linear regression model is formulated in terms of the interval relative frequencies both classical and generalized least square estimators are derived, a non-linear regression model is also formulated in terms of the cumulated relative frequencies both classical and generalized least square estimators are derived. Numerical illustrations are provided in Section 4.

§2. The Model and the MLE for θ

Suppose that in some medical or industrial set-up, inspections occur at times $k\triangle$, $k \geq 1$, where $\triangle$ is some positive constant (known), and that the time a subject enters the study is exactly recorded. Assume without loss of generality that $\triangle = 1$. The observational data consists of independent and identically distributed (i.i.d.) pairs of random variables (U_i, V_i), $i = 1, 2, \ldots, n$, where U_i denotes the time when the subject i enter and V_i the time from entry till the response is observed. Assume without loss of generality that $0 \leq U_i \leq 1$ for all i, $i = 1, 2, \ldots, n$. The true response time is denote by X_i (unknown). It is clear that there is an integer k_i such that $X_i \in (k_i, k_i + 1)$ (so X_i is called the "interval censored data"). Figure 1 helps to clarify the relationship between the variables U, V and X (omit the subscript i for convenient).

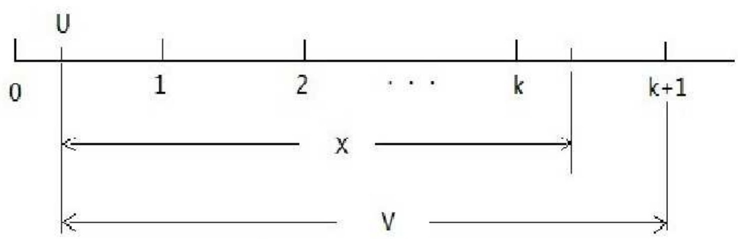

Figure 1 The relationship between U, V and X

Remark 1 From the above description, sample data is composed of (U_i, V_i), $i =$

$1, 2, \ldots, n$, $k_i = (U_i + V_i) - 1$, and other variables in the following discussion are all unknown.

Remark 2 In an interval censoring model, the inspection intervals may be different for each subject, and do not have to be in the same length. However, it is convenience to consider the more special model.

Assume that

$$\mathsf{P}_\theta(U \le u, X \le x) = H(u)F(x;\theta), \tag{2.1}$$

where θ is the unknown parameter, $F(x;\theta)$ is given by (1.2), and $H(u)$ is, in general, an unknown distribution function on $[0,1]$. This assumption can be understood as censoring mechanism has nothing to do with the considering distribution.

By the fact that $V = [U + X] + 1 - U$, and straightforward calculations, where $[\cdot]$ is the greatest integer that is less than or equal to a given real number, we have the CDF of V:

$$\begin{aligned} W(v;\theta) = \mathsf{P}_\theta(V \le v) &= \int_0^1 F([v+u]-u;\theta)\mathrm{d}H(u) \\ &= 1 - \int_0^1 ([v+u]-u)^{-\theta}\mathrm{d}H(u). \end{aligned} \tag{2.2}$$

Differentiating this with respect to v, we see that V has density

$$w(v;\theta) = [F(v;\theta) - F(v-1;\theta)]h([v]+1-v), \tag{2.3}$$

with respect to the Lebesgue measure, where h is density of H.

Taking (1.1) into (2.3), and noticing that $v \ge 2$, we have

$$w(v;\theta) = \begin{cases} \{(v-1)^{-\theta} - v^{-\theta}\}h([v]+1-v) & \text{for } v \ge 2; \\ \{1 - v^{-\theta}\}h(2-v) & \text{for } 1 \le v < 2. \end{cases} \tag{2.4}$$

Now suppose that the observations are reported in the form of (U_i, V_i), $i = 1, 2, \ldots, n$. Following from (1.2), (2.1), (2.2) and (2.3), the probability mass function of observations is given by

$$L(\theta) = \prod_{k=1}^n w(v_{(k)};\theta),$$

it follows from (2.4) that

$$L(\theta) = \prod_{k=1}^{j_0} \{1 - v_{(k)}^{-\theta}\}h(2 - v_{(k)}) \prod_{k=j_0}^n \{(v_{(k)}-1)^{-\theta} - v_{(k)}^{-\theta}\}h([v_{(k)}] + 1 - v_{(k)}),$$

the equation determining the maximum likelihood estimator (MLE) $\widehat{\theta}_1$ for θ is

$$\sum_{k=1}^{j_0}\frac{\ln v_{(k)}}{v_{(k)}^{\theta}-1}+\sum_{k=j_0}^{n}\frac{(v_{(k)}-1)^{-\theta}\ln(v_{(k)}-1)-v_{(k)}^{-\theta}\ln v_{(k)}}{(v_{(k)}-1)^{-\theta}-v_{(k)}^{-\theta}}=0. \tag{2.5}$$

By standard maximum likelihood theory, the estimator $\widehat{\theta}_1$ will be asymptotically unbiased, consistent, efficient, and have asymptotically variance equal to the reciprocal of the information measure.

§3. Regression of Interval Frequencies

Suppose that a set of values, say $1=x_0<x_1<\cdots<x_m<x_{m+1}=\infty$, have been preselected to mark off intervals over the range of $x\in[1,\infty)$. It is easy to obtain n_j, the number of observations falling within each of the intervals: $[x_j,x_{j+1})$, $j=0,1,\ldots,m$, $m<n$. Here, n_j is also called the absolute frequencies, and m is specified in advance according to a practical problem.

Let y_j denote the cumulated relative frequencies of greater-than x_j. That is,

$$y_j=\sum_{t=j}^{m}\frac{n_t}{n},$$

and the probability density function (1.1) implies that

$$\pi_j=\mathsf{P}\{X>x_j\}=x_j^{-\theta},$$

so that $\ln\pi_j=-\theta\ln x_j$, this suggests that we write

$$\ln y_j=-\theta\ln x_j+e_j$$

for some error e_j. Since this model has the appearance of a linear regression of $\ln y_j$ on $\ln x_j$, we may apply ordinary least square to obtain the estimator, say $\widehat{\theta}_2$ for θ. It is easy to see that

$$\widehat{\theta}_2=-\frac{\sum\limits_{j=0}^{m}(\ln y_j)(\ln x_j)}{\sum\limits_{j=0}^{m}(\ln x_j)^2}, \tag{3.1}$$

with variance

$$\mathsf{Var}\,(\widehat{\theta}_2)=\frac{1}{m-1}\sum_{j=0}^{m}(\ln y_j-\widehat{\theta}_2\ln x_j)^2. \tag{3.2}$$

For the sake of the cumulative character of the y_j's, the e_j's homoscedasticity may be ruled out, which may not guarantee that $\widehat{\theta}_2$ has good statistical properties. We now take up a regression for the estimator of θ with the following model (Here, we discard f_0 (and p_0) since $f_0 = 1 - \sum\limits_{j=1}^{m} f_j$ is redundant.):

$$\boldsymbol{f} = \boldsymbol{p} + \boldsymbol{\epsilon}, \tag{3.3}$$

where $\boldsymbol{f} = (f_1, f_2, \ldots, f_m)^T$, $\boldsymbol{p} = (p_1, p_2, \ldots, p_m)^T$, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_m)^T$, and

$$f_j = \frac{n_j}{n}, \qquad p_j = \mathsf{P}(x_j \le X < x_{j+1}) = \int_{x_j}^{x_{j+1}} f(x)\mathrm{d}x. \tag{3.4}$$

From (1.1), (3.4) is simplified as

$$p_j = \begin{cases} x_j^{-\theta} - x_{j+1}^{-\theta} & \text{for } j = 0, 1, \ldots, m-1; \\ x_m^{-\theta} & \text{for } j = m. \end{cases}$$

It is clear that the frequencies $(n_0, n_1, \ldots, n_m)$ comprise a random sample of size n from the multinomial distribution Multi$(n; p_0, p_1, \ldots, p_m)$. Thus, the means, variances, and covariances of the relative frequencies f_j's are:

$$\mathsf{E}(f_j) = p_j, \qquad \mathsf{Var}\,(f_j) = \frac{1}{n} p_j (1 - p_j), \qquad \mathsf{Cov}\,(f_j, f_s) = -\frac{1}{n} p_j p_s \quad (j \ne s). \tag{3.5}$$

Associating (3.3) with (3.5), we have

$$\mathsf{E}(\boldsymbol{\epsilon}) = \mathbf{0}, \qquad \mathsf{Var}\,(\boldsymbol{\epsilon}) = \frac{1}{n}\boldsymbol{\Sigma},$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Lambda} - \boldsymbol{p}\boldsymbol{p}^T = \boldsymbol{\Lambda}(\boldsymbol{I} - \mathbf{1}_m \boldsymbol{p}^T)$ and $\boldsymbol{\Lambda} = \begin{pmatrix} p_1 & 0 & 0 & \cdots & 0 \\ 0 & p_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & p_m \end{pmatrix}$.

To fit (3.3), we first consider a classical least squares approach: choosing θ (and thus $\boldsymbol{p}$) to minimize $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = (\boldsymbol{f} - \boldsymbol{p})^T(\boldsymbol{f} - \boldsymbol{p})$. The least squares estimator $\widehat{\theta}_3$ satisfies the following equation:

$$\widehat{\boldsymbol{p}}_\theta^T (\boldsymbol{f} - \widehat{\boldsymbol{p}}) = 0,$$

where $\widehat{\boldsymbol{p}}$ is the estimator of $\boldsymbol{p}$ and $\boldsymbol{p}_\theta = \partial \boldsymbol{p}/\partial\theta$. Linearizing this equation around the true parameter point θ, i.e. via $\boldsymbol{p} = \widehat{\boldsymbol{p}} + \widehat{\boldsymbol{p}}_\theta(\theta - \widehat{\theta})$, gives

$$\boldsymbol{p}_\theta^T (\boldsymbol{f} - \boldsymbol{p}) - \boldsymbol{p}_\theta^T \boldsymbol{p}_\theta (\widehat{\theta}_3 - \theta) = 0$$

or

$$\widehat{\theta}_3 = \theta + (\boldsymbol{p}_\theta^T \boldsymbol{p}_\theta)^{-1} \boldsymbol{p}_\theta \boldsymbol{\epsilon}.$$

It is clear that $\widehat{\theta}_3$ is asymptotically unbiased and consistent, with asymptotic variance

$$\mathsf{Var}\,(\widehat{\theta}_3) = \frac{1}{n}(\boldsymbol{p}_\theta^T\boldsymbol{p}_\theta)^{-2}\boldsymbol{p}_\theta^T\boldsymbol{\Sigma}\boldsymbol{p}_\theta. \tag{3.6}$$

A generalized least squares estimator $\widehat{\theta}_4$, on the other hand, minimizes the following equation:

$$\begin{aligned} Q &\triangleq \boldsymbol{\epsilon}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\epsilon} \\ &= (\boldsymbol{f}-\boldsymbol{p})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{f}-\boldsymbol{p}) \\ &= (\boldsymbol{f}-\boldsymbol{p})^T\Big(\boldsymbol{\Lambda}^{-1}+\frac{\mathbf{1}_m\mathbf{1}_m^T}{p_0}\Big)(\boldsymbol{f}-\boldsymbol{p}) \\ &= \sum_{j=0}^{m} f_j^2 p_j^{-1} - 1. \end{aligned}$$

Differentiating Q with respect to θ leads to the equation that $\widehat{\theta}_4$ should satisfy:

$$\begin{aligned} \frac{\mathrm{d}Q}{\mathrm{d}\theta} &= -\boldsymbol{p}_\theta^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{f}-\boldsymbol{p}) \\ &= -\sum_{j=0}^{m} f_j^2 p_j^{-2}\frac{\mathrm{d}p_j}{\mathrm{d}\theta}, \end{aligned} \tag{3.7}$$

or equivalently,

$$\sum_{j=0}^{m-1}\Big[\Big(\frac{f_j}{f_m}\Big)^2\cdot\frac{(-x_j^{-\theta}\ln x_j + x_{j+1}^{-\theta}\ln x_{j+1})x_m^{-\theta}}{(x_j^{-\theta}-x_{j+1}^{-\theta})^2\ln x_m}\Big] = 1.$$

$\widehat{\theta}_4$ is asymptotically unbiased and consistent, with asymptotic variance (The asymptotic variance of $\widehat{\theta}_4$ comes from the linearization of (3.7) around the true value θ, a similar procedure that has been used for $\mathsf{Var}\,(\widehat{\theta}_3)$.):

$$\mathsf{Var}\,(\widehat{\theta}_4) = \frac{1}{n}(\boldsymbol{p}_\theta^T\boldsymbol{\Sigma}^{-1}\boldsymbol{p}_\theta)^{-1}. \tag{3.8}$$

Theorem 3.1 Following from (2.5), (3.2), (3.6) and (3.8), there is

$$\mathsf{Var}\,(\widehat{\theta}_1) \le \mathsf{Var}\,(\widehat{\theta}_2), \tag{3.9}$$

$$\mathsf{Var}\,(\widehat{\theta}_1) \le \mathsf{Var}\,(\widehat{\theta}_4), \tag{3.10}$$

$$\mathsf{Var}\,(\widehat{\theta}_4) \le \mathsf{Var}\,(\widehat{\theta}_3). \tag{3.11}$$

Proof It is clear from (2.5) that the MLE $\widehat{\theta}_1$ use whole information from data (U_i, V_i), $i = 1, 2, \ldots, n$, while $\widehat{\theta}_2$ in (3.1) uses only the grouped data and thus only part of the information from data (U_i, V_i), $i = 1, 2, \ldots, n$. This implies (3.9) and (3.10).

From Kantorovich inequality

$$1 \le \frac{\boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} \cdot \boldsymbol{x}^T \boldsymbol{A}^{-1}\boldsymbol{x}}{(\boldsymbol{x}^T\boldsymbol{x})^2} \le \frac{1}{4}\cdot\frac{(\lambda_1+\lambda_n)^2}{\lambda_1\lambda_n},$$

where $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n$ are the eigenvalues of a $n\times n$ symmetric matrix $\boldsymbol{A}$, we have

$$(\boldsymbol{p}_\theta^T\boldsymbol{p}_\theta)^2 \le \boldsymbol{p}_\theta^T\boldsymbol{\Sigma}\boldsymbol{p}_\theta \cdot \boldsymbol{p}_\theta^T\boldsymbol{\Sigma}^{-1}\boldsymbol{p}_\theta$$

or

$$\mathsf{Var}\,(\widehat{\theta}_3) \ge \mathsf{Var}\,(\widehat{\theta}_4),$$

(3.11) is proved. $\square$

As for comparison with $\widehat{\theta}_2$, and $\widehat{\theta}_4$, it does not appear possible to make a general assertion as to which will have smaller asymptotic variance. It seems, from numerical illustration, that $\widehat{\theta}_4$ has smaller asymptotic variance than $\widehat{\theta}_2$.

§4. Numerical Illustrations

For Pareto distribution with parameters $\theta = 0.5, 1.0, 1.5$ and 2.0, we consider random sample of size $n = 300$, and $\Delta = 1$. Suppose that sample data will be reported in the form of interval censoring, and their relative frequency distribution in 6 intervals, marked off by $x_0 = 1$, $x_1 = 2$, $x_3 = 4$, $x_4 = 8$, $x_5 = 16$, $x_6 = 32$, $x_7 = 64$ and $x_8 = \infty$. Then using the formulas developed in this paper, using Matlab, it is easy to calculate $\widehat{\theta}_1$ and $\widehat{\theta}_4$, and their variance, as for $\widehat{\theta}_2$ and $\widehat{\theta}_3$, can be calculated directly, and their asymptotic variances. The results are presented in Table 1.

Table 1　Asymptotic variances of alternative estimators

	$\theta = 0.5$	$\theta = 1.0$	$\theta = 1.5$	$\theta = 2.0$
$\widehat{\theta}_1$	0.446	1.235	2.635	4.436
$\widehat{\theta}_2$	0.501	1.732	3.563	6.154
$\widehat{\theta}_3$	0.485	1.481	3.112	5.674
$\widehat{\theta}_4$	0.473	1.423	2.874	4.665

From the above simulation results, we can see that, $\widehat{\theta}_1$ has smaller variance, $\widehat{\theta}_3$ has lager variance, and $\widehat{\theta}_4$ has smaller asymptotic variance than $\widehat{\theta}_2$.

§5. Conclusion

If interval censored observations is indeed comes from Pareto distribution, it comes from our discussion that $\widehat{\theta}_1$ is the best estimation in the four kinds of estimators under the variance criterion. In addition, we can't distinguish $\widehat{\theta}_2$ with $\widehat{\theta}_4$, and expect further discussion.

References

[1] Ding, B.J. and Zheng, Z.K., The rates of convergence of the estimation of two-point distribution with interval censored data, *Chinese Journal of Applied Probability and Statistics*, **23(3)**(2007), 292–302.

[2] Ding, B.J., Regression model with an interval-censored data covariant, *Chinese Journal of Applied Probability and Statistics*, **28(2)**(2012), 143–149.

[3] Ding, B.J., The estimation of a distribution with interval censored data and its convergence rate, *Journal of Systems Science and Mathematical Sciences*, **28(6)**(2008), 641–648.

[4] Lee, C., An urn model in the simulation of interval censored failure time data, *Statistics & Probability Letters*, **45(2)**(1999), 131–139.

[5] Akritas, M.G., Rank tests with interval-censored data, *The Annals of Statistics*, **16(4)**(1988), 1490–1502.

[6] Gentleman, R. and Geyer, C.J., Maximum likelihood for interval censored data: consistency and computation, *Biometrika*, **81(3)**(1994), 618–623.

[7] Topp, R. and Gomez, G., Residual analysis in linear regression models with an interval-censored covariate, *Statistics in Medicine*, **23(21)**(2004), 3377–3391.

[8] Turnbull, B.W., The empirical distribution function with arbitrarily grouped, censored and truncated data, *Journal of the Royal Statistical Society, Series B* (*Methodological*), **38(3)**(1976), 290–295.

[9] Chang, M.N. and Yang, G.L., Strong consistency of a nonparametric estimator of the survival function with doubly censored data, *The Annals of Statistics*, **15(4)**(1987), 1536–1547.

区间删失情况下的Pareto分布估计

丁 邦 俊

(喀什师范学院, 喀什, 844006; 华东师范大学金融与统计学院, 上海, 200241)

考虑观察数据为区间删失情况下的Pareto分布参数估计问题, 利用最大似然估计方法和最小二乘估计方法给出了参数的四个估计量, 比较了它们的方差, 最后给出了模拟报告.

关键词: Pareto分布, 区间删失数据, 参数估计.

学科分类号: O213.7.