ESTIMATION OF PARAMETERS FOR PARETO DISTRIBUTION

Eldesoky E. Afify.

Faculty of Eng. Dept. of Mathematics, Minoufiya Univ. Shibeen El Kom, Egypt.

Keywords: Pareto distribution, least squares method, ridge regression, maximum product of spacing, outliers, root mean squared errors.

Abstract

In This paper, we use least squares, ridge regression and maximum product of spacing methods to estimate the parameters for Pareto distribution. Root mean squared errors are used to compare between these methods in the presence of outliers. Numerical examples are worked out.

1. Introduction

Pareto distribution has a wide use in economic studies. It has played a major part in investigations of several economic phenomena. Recently it has been used to the study of ozone levels in the upper atmosphere, tinsile strength of nylon carpet fibers. It has played a very important role in the investigation of city population occurrence of natural resources, insurance risk, business failures. Wingo (1982) discussed the unimodality of the conditional likelihood function of the present distribution using multi censored samples. He concluded that practical one since its presence facilitates as a computational search for the maximum likelihood estimates. Arnold and Press (1983) gave an extensive historical survey

of its use in the context of income distribution. Samia and Mohamed (1993) used five modifications of moments to estimate the parameters of the Pareto distribution. Kang and Young (1997) estimated the parameters of a Pareto distribution by Jackknife and bootstrap methods. Abdel Ghaly, Attia and Aly (1998) obtained the prediction of the shape parameter of Pareto distribution.

In this paper, we estimate the parameters of the Pareto distribution by least squares, ridge regression and maximum product of spacing methods in the presence of outliers. We compare between these methods by using root mean squared errors. Two parameters Pareto distribution has a probability density function given by

$$f(t; c,k) = \frac{c}{t}\left(\frac{k}{t}\right)^{c} \quad t> k,\ c,\ k > 0 \qquad (1)$$

Where c and k are the parameters of the distribution. The cumulative distribution function and reliability function are given respectively by

$$F(t; c, k) = 1 - \left(\frac{k}{t}\right)^{c} \qquad (2)$$

$$R(t) = (t/k)^{-c} \qquad (3)$$

2. Least squares (LS):

Taking logarithm of both sides of (3) we get

$\text{Log } R(t_i) = -c \text{ Log } t_i + c \text{ Log } k$, $\quad i=1,2,\ldots,n$. Then

$$\text{Log } t_i = \text{Log } k - \frac{1}{c} \text{ Log } R(t_i) \qquad (4)$$

Equation (4) can be written in the form

$$Y_i = A + B\,X_i \qquad (5)$$

Where $A = \text{Log } k$, $B = -1/c$, $Y_i = \text{Log } t_i$ and $X_i = \text{Log } R\,(t_i)$

$$\text{Let} \qquad S = \sum_{i=1}^{n} (Y_i - A - BX_i)^2 \qquad (6)$$

Differentiation S w.r.t. A and B then equate to zero, we have the least square (LS) estimates of A and B.

$$\hat{A} = \frac{\sum Log\,R(t_i) \sum Log\,R(t_i)\,Log(t_i) - \sum (Log\,R(t_i))^2 \sum Log\,t_i}{\left(\sum Log\,R(t_i)\right)^2 - n \sum (Log\,R(t_i))^2} \qquad (7)$$

$$\hat{B} = \frac{\sum Log\,R(t_i) \sum Log\ t_i - n \sum Log\,R(t_i)\,Log\,t_i}{\left(\sum Log\,R(t_i)\right)^2 - n \sum (Log\,R(t_i))^2} \qquad (8)$$

3. Ridge Regression (RR):

The ridge regression (RR) estimates of A and B can be obtained by minimizing the error sum of squares for the model (5) subject to the single constraint that $A^2 + B^2 = \rho$, where ρ is a finite positive constant. The method of Lagrange multipliers requires the differentiation of

$$L = \sum_{i=1}^{n} (Y_i - A - BX_i)^2 + \lambda(A^2 + B^2 - \rho)$$

With respect to A and B. When these derivatives are equated to zero, we obtain the following two equations

$$\sum_{i=1}^{n} Y_i = (n + \lambda)A + B \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i Y_i = A\sum_{i=1}^{n} X_i + B(\lambda + \sum_{i=1}^{n} X^2{}_i)$$

Solving the last two equations for A and B, we obtain the estimates of A and B as

$$\hat{A} = \frac{\sum Log\, R(t) \sum Log\, R(t_i)\, Log(t_i) - \sum (Log\, R(t_i))^2 \sum (\lambda + \sum Log\, t_i)}{(\sum Log\, R(t_i))^2 - (n+\lambda)(\lambda + \sum (Log\, R(t_i))^2)} \qquad (9)$$

$$\hat{B} = \frac{\sum Log\, R(t) \sum Log\, t_i - (n+\lambda) \sum Log\, R(t_i)\, Log\, t_i}{(\sum Log\, R(t_i))^2 - (n+\lambda)(\lambda + \sum (Log\, R(t_i))^2)} \qquad (10)$$

Where O < λ < 1 is the ridge coefficient. The readers may see Ronald and Raymond (1978). If λ=o, we obtain the least square estimates.

4. Maximum product of spacing (MPS):

We now discuss the MPS method introduced by Cheng and Amin (1983) that gives consistent estimators under much more general conditions it gives also efficient estimators as shown by Cheng and Amin (1983) and Ranneby (1984) independently. The estimates by this method are obtained by maximizing the geometric mean of the spacing.

$$G(c, k) = \frac{1}{n+1} \sum_{i=1}^{n+1} Log\ (F(t_i; c, k) - F(t_{i-1}; c, k))$$

$$= \frac{1}{n+1} \sum_{i=1}^{n+1} Log \left[\left(\frac{k}{t_{i-1}} \right)^c - \left(\frac{k}{t_i} \right)^c \right] \qquad (11)$$

The partial derivative of (11) w.r.t. c and k give the following two MPS simultaneous equations:

$$(n+1)\frac{\partial G}{\partial c}=\sum_{i=1}^{n+1}\frac{\left(\frac{k}{t_{i-1}}\right)^{c}Log\left(\frac{k}{t_{i-1}}\right)-\left(\frac{k}{t_{i}}\right)^{c}Log\left(\frac{k}{t_{i}}\right)}{\left(\frac{k}{t_{i-1}}\right)^{c}-\left(\frac{k}{t_{i}}\right)^{c}}=0 \qquad (12)$$

$$(n+1)\frac{\partial G}{\partial k}=\sum_{i=1}^{n+1}\frac{ck^{c-1}/t_{i-1}^{c}-ck^{c-1}/t_{i}^{c}}{\left(\frac{k}{t_{i-1}}\right)^{c}-\left(\frac{k}{t_{i}}\right)^{c}}=0 \qquad (13)$$

To obtain the estimates $\hat{c}$ *and* $\hat{k}$, we solved Equations (12) and (13) by using Newton Raphson method.

5. Mean Shift Outlier:

Outliers are observations which appear to be inconsistent with the rest of the observations. Outlier theory is quite popular and many studies on outliers are based on the mean shift model. The presence of an outlier at the ith data point produces a location shift. Readers interested in outliers may see Tietjen, Moore and Beckman (1973). Outliers in our study were generated from

The mean shift model, $y= x\beta +\varphi\delta_i +\varepsilon$

where δi is an nX1 column vector with a one in the ith row and zero elsewhere and φ is the regression coefficient of the unit vector δi. Under the null hypothesis of no outlier H_o: $E(y)=x\beta$ and H_1: $y= x\beta +\varphi\delta_i$. For each n, one or two observations were randomly selected and an outlier(s) created by adding φ to the model. The values of φ are chosen to be $1\sigma(1\sigma)3\ \sigma$.

6. Simulation Study:

To assess the performance of these methods, the estimates and the root mean squared errors (RMSE) for each method were calculated using 10,000 replications for each sample size. In our simulation study we generate a data set for certain values of c and k and for sample sizes of 10, 20 and 30. The true value pairs (c,k): (1,1), (1,2), (2,1) and (3,3). The generation of random sample by observing that if U is uniform (0,1), then $X = k(1 - U)^{-1/c}$ is Pareto (c, k).

Table (1) Estimates of parameters with no outliers, n=10

N	Method	c, k	$\hat{c}$	$\hat{k}$	RMSE
10	LS	1,1	1.001	0.991	$261x10^{-6}$
		1,2	1	1.9997	$242x10^{-6}$
		2,1	1.99953	0.9997	$195x10^{-6}$
		3,3	2.99787	2.99956	$214x10^{-6}$
	RR	1,1	1.014	0.992	$266x10^{-6}$
		1,2	1.00008	1.999979	$898x10^{-3}$
		2,1	2.16	1.042	$48x10^{-3}$
		3,3	0.874	1.15	$312x10^{-3}$
	MPS	1,1	1.012	0.987	$269x10^{-6}$
		1,2	1.0007	1.9886	$35x10^{-4}$
		2,1	1.931	0.98	$199.7x10^{-4}$
		3,3	0.997	1.293	$314x10^{-3}$

Table (2) Estimates of parameters with no outliers, n=20

n	Method	c, k	$\hat{c}$	$\hat{k}$	RMSE
20	LS	1,1	1	1.9997	$59x10^{-6}$
		1,2	0.99917	2.00189	$175x10^{-8}$
		2,1	2.005	0.99879	$197x10^{-6}$
		3,3	2.9997	3.0002	$581x10^{-5}$
	RR	1,1	1.00008	1.9998	$585x10^{-3}$
		1,2	1.00008	1.99979	$7526x10^{-7}$
		2,1	1.901	0.9474	$514x10^{-4}$
		3,3	0.99	1.135	$325x10^{-3}$
	MPS	1,1	1.021	1.013	$543x10^{-4}$
		1,2	1.0007	1.9886	$6632x10^{-5}$
		2,1	1.936	1.043	$584x10^{-4}$
		3,3	0.893	1.568	$217x10^{-4}$

Table(3) Estimates of parameters with no outliers, n=30

n	Method	c, k	$\hat{c}$	$\hat{k}$	RMSE
30	LS	1,1	1.36974	1.2025	$123x10^{-6}$
		1,2	1.00014	2.00024	$551x10^{-6}$
		2,1	1.99998	1.000012	$157x10^{-7}$
		3,3	2.9998	2.99977	$118x10^{-7}$
	RR	1,1	1.2092	1.025	$429x10^{-6}$
		1,2	0.711	1.043	$193x10^{-5}$
		2,1	2.032	1.012	$1109x10^{-5}$
		3,3	0.951	1.045	$3067x10^{-5}$
	MPS	1,1	1.311	1.116	$687x10^{-5}$
		1,2	0.813	1.146	$197x10^{-4}$
		2,1	2.114	1.018	$148x10^{-4}$
		3,3	1.965	1.179	$464x10^{-4}$

Table (4) Estimates of parameters with outliers, n=10

n	Method	c, k	$\hat{c}$	$\hat{k}$	RMSE
10	LS	1,1	1.1655	1.002	$1189x10^{-5}$
		1,2	1.2209	2.501	$143x10^{-5}$
		2,1	3.2765	1.2713	$603x10^{-5}$
		3,3	3.904	3.7874	$843x10^{-5}$
	RR	1,1	1.165	1.00215	$1189x10^{-4}$
		1,2	1.1794	2.203	$459x10^{-4}$
		2,1	2.137	1.059	$722x10^{-4}$
		3,3	3.205	2.57	$258x10^{-4}$
	MPS	1,1	1.1617	1.00216	$118x10^{-4}$
		1,2	1.1781	1.989	$472x10^{-4}$
		2,1	2.001	1.006	$763x10^{-4}$
		3,3	3.216	2.765	$158x10^{-4}$

Table (5) Estimates of parameters with outliers, n=20

n	Method	c, k	$\hat{c}$	$\hat{k}$	RMSE
20	LS	1,1	1.012	1.029	$149x10^{-5}$
		1,2	1.0616	2.0238	$129x10^{-5}$
		2,1	1.9619	1.038	$5x10^{-5}$
		3,3	2.991	3.011	$726x10^{-5}$
	RR	1,1	0.9981	1.003	$225x10^{-4}$
		1,2	1.062	2.02291	$129x10^{-4}$
		2,1	1.795	1.0534	$848x10^{-4}$
		3,3	1.3152	1.0942	$415x10^{-3}$
	MPS	1,1	1.072	0.9975	$2008x10^{-5}$
		1,2	1.064	2.02292	$129x10^{-5}$
		2,1	2.114	0.9961	$188x10^{-4}$
		3,3	2.753	2.761	$1124x10^{-4}$

Table (6) Estimates of parameters with outliers, n=30

n	Method	c, k	$\hat{c}$	$\hat{k}$	RMSE
30	LS	1,1	1.464	1.303	$205x10^{-3}$
		1,2	1.03384	2.132	$354x10^{-4}$
		2,1	1.03384	1.0873	$2289x10^{-4}$
		3,3	3.5296	3.2368	$1366x10^{-4}$
	RR	1,1	1.20773	1.02736	$419x10^{-4}$
		1,2	0.6939	1.0429	$1861x10^{-4}$
		2,1	2.02	1.013	$1309x10^{-5}$
		3,3	1.84	2.86	$963x10^{-4}$
	MPS	1,1	0.9983	0.99986	$415x10^{-6}$
		1,2	1.003	2.008	$237x10^{-5}$
		2,1	2.002	0.9998	$247x10^{-6}$
		3,3	2.7611	2.8613	$569x10^{-4}$

7- Conclusions:

Least squares method is easy to be computed and takes little compute time. The RR and MPS take more computer time than least squares. The RMSE from RR and MPS methods is larger than RMSE from LS method .It has been shown from computational results that the estimators are much closer to the true parameter values when no outliers were present.

References

1- Abdel Ghaly, A.A, Attia, A.F. and Ali, H.M. (1998). Estimation of the parameters of Pareto distribution and the reliability function using accelerated life testing with censoring. Commun. Statist., Simul A. 27 (2), 469-484.

2- Arnold, B. C. and Press, S. J. (1983) Bayesian inference for Pareto populations. Journal of Econometrics, 21, 287-306.

3- Cheng, G. C. H. and Amin, N. A. K. (1983) Estimating parameters in continuous univariate distributions with as shifted origin. J. R. Statist. Soc. B. 45 (3), 394-403.

4- Kang, S. B. and Young, C. S.(1997) Estimation of the parameters in a Pareto distribution by jackknife and bootstrap methods.J. Information and Optimization Sci.18, 289-300.

5- Ranneby, Bo (1984) The maximum spacing method. An estimation method related to the maximum likelihood method. Scand. J. Statist., 11, 93-112.

6- Ronald, E.W. and Raymond, H.M. (1978) Probability and statistics for engineers and scientists. Second Edition. Macmillan publishing Co., Inc, New York.

7- Samia, A.S. and Mohamed, M. (1993). Modified moment estimators for the 3 parameter Pareto distribution. The annual conference, ISSR, Cairo Univ. vol. (28), part (2).

8- Tietjen, G.L., Moore, R.L. and Beckman, R.J.(1973) Testing for a single outlier in simple linear regression, Technometrics, 15,717-721.

9- Wingo, D. R. (1982) Unimodality of the Pareto distribution likelihood function for multi censored samples and implications for estimation. Commun. Statistics, Theory and Methods. 11 (10), 1129- 1138.