RELIK 2011; Praha, 5. a 6. 12. 2011

THREE-PARAMETRIC LOGNORMAL DISTRIBUTION AND ESTIMATING ITS PARAMETERS USING THE METHOD OF L-MOMENTS

Diana Bílková

Abstract

Common statistical methodology for description of the statistical samples is based on using conventional moments or cummulants. An alternative approach is based on using different characteristics which are called the L-moments. The L-moments are an analogy to the conventional moments, but they are based on linear combinations of the order statistics, i.e. L-statistics. Parameter estimations using the L-moments are especially in the case of small samples often more precise than estimates calculated using the maximum likelihood method. This text concerns with the application of the L-moments in the case of larger samples and with the comparison of the precision of the method of L-moments with the precision of other methods (moment, quantile and maximum likelihood method) of parameter estimatimation in the case of larger samples. We used two types of data, namely data sets of individual data and data ordered to the form of interval frequency distribution. Three-parametric lognormal distribution was used as the theoretical model.

Key words: Lognormal distribution, L-moments, L-moment method of parameter estimation, income distribution, wage distribution

JEL Code: C13, C16

Introduction

From the statistical literature it is well-known the use of L-moments in connection with the data from the field of hydrology and meteorology (for example rainfall). In such cases, there are generally relatively small data sets. This paper deals with the use of L-moments in the case of large data sets. There are the data of two types, namely, individual data on net annual household income per capita in CZK (years 1992, 1996 and 2002 – statistical survey Microcensus and years 2005, 2006, 2007 and 2008 – statistical survey SILC), and second, data sorted into a form of interval frequency distribution, these data refer to gross monthly wage in CZK (from official web page of the Czech Statistical Office). In both cases we compare the accuracy of the method of L-moments with an accuracy of other methods of parameter estimation. Income data come from the statistical surveys Microcensus and SILC of the Czech Statistical Office, while the wage data come from official website of the Czech Statistical Office. Three-parametric lognormal distribution was used as the basic parametric distribution. Accuracy of the method of L-moments was compared with the accuracy of other methods of parameter estimation, such as moment method, quantile method, maximum likelihood method.

The question of income distribution is in statistical literature rather extensively treated, see for example Bartošová (2006) or Bílková (2008). The methods of estimation parameters, including the three-parametric lognormal distribution are described in the statistical literature, see for example in Bílková (2008). Three-parametric lognormal distribution is discussed in detail for example in Bartošová, Bína (2009) or in Bílková (2008), moment method of parameter estimation in Bílková (2008), quantile method in Bílková (2008) or Sipková, Sodomová (2009), maximum likelihood method in Bílková (2008). For example paper Miskolczi, Langhamrová (2011) is based inter alia on the knowledge of income distribution. The concept of L-moments and the use of these quantities in the estimation of parameters of probability distribution are given in Bílková (2011), in Hosking (1990) or in Hosking, Wales (1997).

1 Methodology

The essence of lognormal distribution is treated in detail for example in Aitchison and Brown (1957). Use of lognormal distribution in connection with wage or income distributions is described in Bartošová (2006) or Bílková (2008). The methods of parameter estimation (moment method, quantile method, maximum likelihood method) are described in detail for example in Bílková (2008).

1.1 L-Moment Method of Parameter Estimation

Question of L-moment is described in detail for example in Hosking and Wales (1997). We will assume that X is a real random variable with the distribution function $F(x)$ and quantile function $x(F)$ and $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$ are the order statistics of the random sample of the size n selected from the distribution X. Then the r-th L-moment of the random variable X is defined as

$$\lambda_r = r^{-1} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} EX_{r-k:r}, \quad r = 1, 2, 3, \ldots. \tag{1}$$

The letter ‘L’ in the name ‘L-moments’ is to stress the fact that r-th L-moment λ_r is a linear function of the expected order statistics. Natural estimate of the L-moment λ_r based on the observed sample is furthermore a linear combination of the ordered values, i.e. the so called L-statistics. The expected value of the order statistic is of the form

$$EX_{j:r} = \frac{r!}{(j-1)! \cdot (r-j)!} \cdot \int x\,[F(x)]^{j-1} \cdot [1-F(x)]^{r-j}\,\mathrm{d}F(x). \tag{2}$$

The first four L-moments are of the form

$$\lambda_1 = EX = \int_0^1 x(F)\,\mathrm{d}F, \tag{3}$$

$$\lambda_2 = \frac{1}{2} E(X_{2:2} - X_{1:2}) = \int_0^1 x(F) \cdot (2F - 1)\,\mathrm{d}F, \tag{4}$$

$$\lambda_3 = \frac{1}{3} E(X_{3:3} - 2X_{2:3} + X_{1:3}) = \int_0^1 x(F) \cdot (6F^2 - 6F + 1)\,\mathrm{d}F, \tag{5}$$

$$\lambda_4 = \frac{1}{4} E(X_{4:4} - 3X_{3:4} + 3X_{2:4} - X_{1:4}) = \int_0^1 x(F) \cdot (20F^3 - 30F^2 + 12F - 1)\,\mathrm{d}F. \tag{6}$$

Details about the L-moments can be found in Guttman (1993) or Hosking (1990). We get the first three L-moments of the three-parametric lognormal distribution LN(μ, σ^2, ξ), which is described e.g. in Hosking (1990). The following relations are valid for these L-moments

$$\lambda_1 = \xi + \exp\left(\mu + \frac{\sigma^2}{2}\right), \tag{7}$$

$$\lambda_2 = \exp\left(\mu + \frac{\sigma^2}{2}\right) \cdot \operatorname{erf}\left(\frac{\sigma}{2}\right), \tag{8}$$

$$\tau_3 = \frac{6\pi^{-1/2}}{\operatorname{erf}\left(\frac{\sigma}{2}\right)} \cdot \int_0^{\sigma/2} \operatorname{erf}\left(\frac{x}{\sqrt{3}}\right) \cdot \exp(-x^2)\,\mathrm{d}x, \tag{9}$$

where erf(z) is the so called error function defined as

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \cdot \int_0^z e^{-t^2}\,\mathrm{d}t. \tag{10}$$

Now we will assume that $x_1, x_2, \ldots, x_n$ is a random sample and $x_{1:n} \le x_{2:n} \le \ldots \le x_{n:n}$ is the ordered sample. The r-th sample L-moment is defined as

$$l_r = \binom{n}{r}^{-1} \cdot \sum_{1 \le i_1 \le i_2 \le \ldots \le i_r \le n} \sum \ldots \sum r^{-1} \cdot \sum_{k=0}^{r-1} (-1)^k \cdot \binom{r-1}{k} \cdot x_{i_{r-k}:n}, \quad r = 1, 2, \ldots, n. \tag{11}$$

We can write specifically for the first four sample L-moments

$$l_1 = n^{-1} \cdot \sum_i x_i, \tag{12}$$

$$l_2 = \frac{1}{2} \cdot \binom{n}{2}^{-1} \cdot \sum_{i>j} \sum (x_{i:n} - x_{j:n}), \tag{13}$$

$$l_3 = \frac{1}{3} \cdot \binom{n}{3}^{-1} \cdot \sum_{i>j>k} \sum \sum (x_{i:n} - 2x_{j:n} + x_{k:n}), \tag{14}$$

$$l_4 = \frac{1}{4} \cdot \binom{n}{4}^{-1} \cdot \sum_{i>j>k>l} \sum \sum \sum (x_{i:n} - 3x_{j:n} + 3x_{k:n} - x_{l:n}). \tag{15}$$

Let us denote the distribution function of the standard normal distribution as Φ, then Φ^{-1} represents the quantile function of the standard normal distribution. The following relation holds for the distribution function of the three-parametric lognormal distribution LN(μ, σ^2, ξ)

$$F = \Phi\left[\frac{\ln(x-\xi)-\mu}{\sigma}\right]. \tag{16}$$

The estimates of the three-parametric lognormal distribution can then be calculated as

$$z = \sqrt{\frac{8}{3}}\cdot\Phi^{-1}\left(\frac{1+t_3}{2}\right), \tag{17}$$

$$\hat{\sigma} \approx 0{,}999\,281\,z - 0{,}006\,118\,z^3 + 0{,}000\,127\,z^5, \tag{18}$$

$$\hat{\mu} = \ln\left[\frac{l_2}{\operatorname{erf}\left(\frac{\hat{\sigma}}{2}\right)}\right] - \frac{\hat{\sigma}^2}{2}, \tag{19}$$

$$\hat{\theta} = l_1 - \exp\left(\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right). \tag{20}$$

More on L-moments is for example in Kyselý and Picek (2007).

1.2 Appropriateness of the Model

In assessing the appropriateness of the constructed model we need to use any of the criterions, which may be for example the sum of all absolute deviations of the observed and theoretical frequencies *S*, eventually known criterion χ^2. The question of the appropriateness of the curve as a model of the income or wage distribution in these large sample sizes, such are in the case of the income and wage distributions encountered, is explained for example in (Bílková, 2008). Graph representing the development of the sample median and of the median of a theoretical distribution using the concrete method of parameter estimation, may bring some insight in terms of accuracy of the method of parameter estimation, too.

2 Analysis and Results

Tabs. 1 to 3 present the estimated parameters of three-parametric lognormal curves using two various methods of point estimation of parameters (method of L-moments and maximum likelihood method) and the sample L-moments on the basis of them the parameters were estimated. We can see from Tab. 3 that the value of the parameter θ (theoretical beginning

Tab. 1: Sample L-moments and parameter estimations of three-parametric lognormal distribution obtained using the L-moment method − Income

	Sample L-moments			Parameter estimation		
Year	l_1	l_2	l_3	**μ**	σ^2	**θ**
1992	35,246.51	7,874.26	2,622.14	9.696	0.490	14,491.687
1996	66,121.92	16,237.54	5,685.46	10.343	0.545	25,362.753
2002	105,029.89	27,978.40	10,229.62	10.819	0.598	37,685.637
2005	111,023.71	28,340.18	9,113.57	11.028	0.455	33,738.911
2006	114,945.08	28,800.68	9,286.18	11.040	0.458	36,606.903
2007	123,806.49	30,126.11	9,530.57	11.112	0.440	40,327.610
2008	132,877.19	31,078.96	9,702.45	11.163	0.428	45,634.578

Source: Own research

Tab. 2: Sample L-moments and parameter estimations of three-parametric lognormal distribution obtained using the L-moment method − Wage

	Sample L-moments			Parameter estimation		
Year	l_1	l_2	l_3	**μ**	σ^2	**θ**
2002	17,437.49	4,251.48	1,267.44	9.238	0.388	4,952.259
2003	18,663.18	4,524.95	1,251.90	9.402	0.332	4,364.869
2004	19,697.57	5,001.34	1,586.09	9.313	0.442	5,872.138
2005	20,738.14	5,262.93	1,636.67	9.392	0.424	5,908.390
2006	21,803.28	5,454.74	1,738.23	9.393	0.447	6,795.207
2007	23,882.83	6,577.65	2,627.93	9.222	0.724	9,349.280
2008	25,477.59	6,993.72	2,737.94	9.319	0.693	9,719.297

Source: Own research

Tab. 3: Parameter estimations of three-parametric lognormal distribution obtained using the maximum likelihood method

Income				Wage			
Year	**μ**	σ^2	**θ**	**Year**	**μ**	σ^2	**θ**
1992	10.384	0.152	-0.342	**2002**	8.977	0.828	6,364.635
1996	10.995	0.180	52.236	**2003**	9.024	0.615	6,679.910
2002	11.438	0.211	73.525	**2004**	9.363	0.306	3,090.038
2005	11.503	0.206	-2.050	**2005**	9.400	0.329	4,134.624
2006	11.542	0.199	-8.805	**2006**	9.159	0.742	8,070.167
2007	11.623	0.190	-42.288	**2007**	9.487	0.369	2,586.616
2008	11.703	0.177	-171.167	**2008**	9.593	0.341	3,324.455

Source: Own research

Tab. 4: Sum of absolute deviations of the observed and theoretical frequencies for all intervals − net annual household income per capita

Year	Method			
	L-moment	**Moment**	**Quantile**	**Maximum likelihood**
1992	2,661.636	5,256.970	3,880.846	2,933.275
1996	5,996.435	15,673.846	9,677.446	7,181.322
2002	2,181.635	3,888.523	3,206.585	2,236.348
2005	1,158.556	2,261.200	1,331.944	1,237.170
2006	2,197.016	3,375.662	2,984.503	2,217.975
2007	2,359.258	3,654.637	2,995.680	2,585.448
2008	2,251.531	4,282.314	3,277.620	2,889.890

Source: Own research

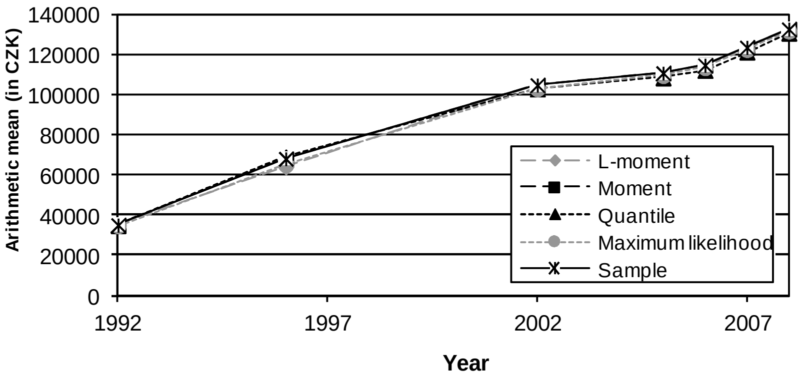
Tab. 5: Sum of absolute deviations of the observed and theoretical frequencies for all intervals − gross monthly wage

Year	Method			
	L-moment	**Moment**	**Quantile**	**Maximum likelihood**
2002	134,846.633	314,497.134	292,479.483	289,279.267
2003	135,772.928	356,423.157	303,335.493	283,469.483
2004	252,042.801	357,087.483	335,019.202	295,900.939
2005	260,527.847	426,062.444	345,954.758	306,785.789
2006	277,661.535	448,632.374	372,420.681	357,828.202
2007	229,525.420	432,745.341	338,552.122	250,114.480
2008	255,510.389	441,371.539	372,924.579	289,621.287

Source: Own research

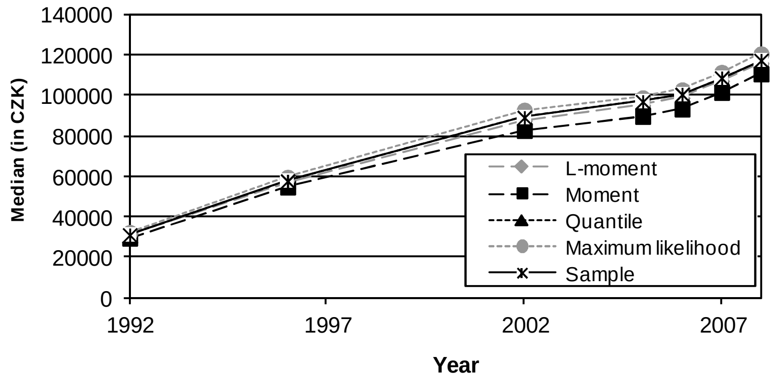
of the distribution) is in some cases negative. This means that lognormal curve gets into negative territory at the beginning of its course. Because of a very tight contact of the lower tail of the lognormal curve with the horizontal axes, this fact does not have to be a problemfor a good fit of the model. The advantage of the lognormal models is that the parameters have an easy interpretation. Also some parametric functions of these models have straight interpretation. Because the estimated value of this parameters is negative, we can not really interpret this value.

Fig. 1: Development of sample average net annual income per capita and the theoretical expected value (in CZK)



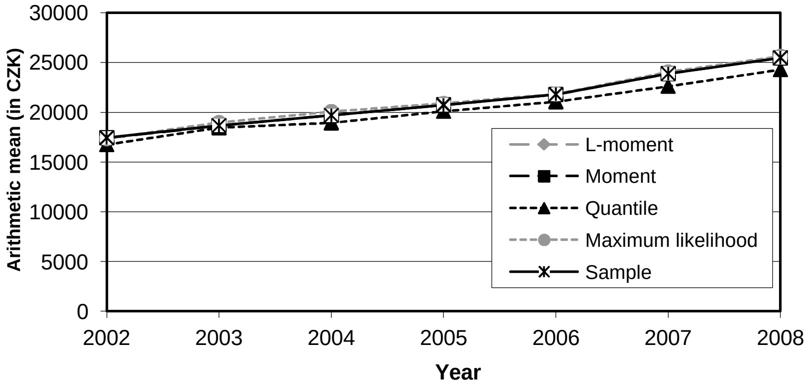
Source: Own research

Fig. 2: Development of sample median of net annual income per capita and the theoretical median (in CZK)



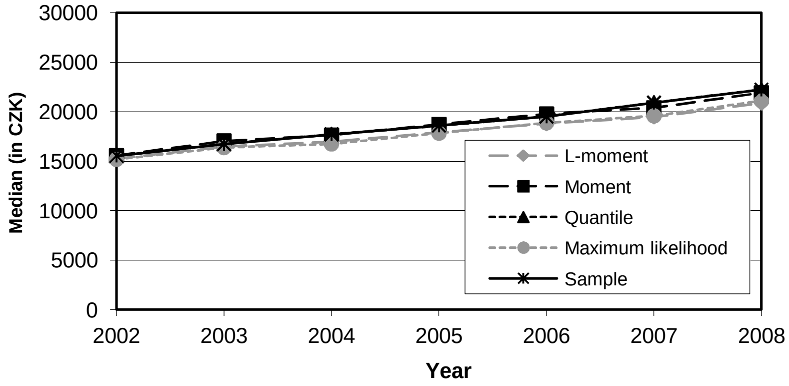
Source: Own research

Fig. 3: Development of sample average gross monthly wage and the theoretical expected value (in CZK)



Source: Own research

Fig. 4: Development of sample median of gross monthly wage and the theoretical median (in CZK)



Source: Own research

Conclusion

Tabs. 4 and 5 provide more accurate information about the used methods of parameter estimation. These tables contain the sum of absolute deviations of the observed and theoretical frequencies for all intervals and therefore they serve as an objective criterion for evaluating the accuracy of used methods of parameter estimation. It should be noted here that in the case of income distribution on the one hand, and in the case of wage distribution on the other hand, we used the same number of intervals, whose width is expanded in time due to the rising level of the distributions. As can be seen from Tabs. 4 and 5, the method of L-moments provides the most accurate results, which are even more accurate than results obtained using the maximum likelihood method. Already mentioned maximum likelihood method ended in terms of accuracy of the estimations as the second best. Quantile method of parameter estimation follows as the third best (second worst). As expected, moment method of parameter estimation provides the least accurate results.

Figures 1 to 4 also provide approximate information about the accuracy of the used methods of parameter estimation. Figures 1 and 3 represent the development of the sample arithmetic mean and the development of theoretical expected values of three-parametric lognormal distribution with parameters estimated using different methods of parameter estimation. Figures 2 and 4 represent the development of the sample median and the development of theoretical medians of three-parametric lognormal distribution with parameters estimated using different methods of parameter estimation.

Acknowledgment

The paper was supported by grant project IGS 24/2010 called “Analysis of the Development of Income Distribution in the Czech Republic since 1990 to the Financial Crisis and Comparison of This Development with the Development of the Income Distribution in Times of Financial Crisis − According to Sociological Groups, Gender, Age, Education, Profession Field and Region” from the University of Economics in Prague.

References

Aitchison, J., Brown, J.A.C. (1957). *The Lognormal Distribution with Special Reference to Its Uses in Economics*. Cambridge University Press, Cambridge.

Bartošová, J. (2006). Logarithmic-Normal Model of Income Distribution in the Czech Republic. *Austrian Journal of Statistics*, Vol. 35, Iss. 23, pp. 215 − 222. ISSN 1026-597x.

Bartošová, J., Bína V. (2009) Modelling of Income Distribution of Czech Households in Years 1996 – 2005. *Acta Oeconomica Pragensia*. Vol. 17, No. 4, pp. 3 – 18. ISSN 0572-3043.

Bílková, D. (2008). Application of Lognormal Curves in Modeling of Wage Distributions. *Journal of Applied Mathematics*, Vol. 1, Iss. 2, pp. 341 − 352. ISSN 1337-6365.

Bílková, D. (2011). Use of the L-Moments Method in Modeling the Wage Distribution. Bratislava February 1 – 4, 2011, Proceedings. In: 10th International Conference *APLIMAT 2011*, Bratislava: Slovak University of Technology, Bratislava, pages 1471 – 1481. ISBN 978-80-89313-52-5.

Guttman, N. B. (1993). The Use of L-moments in the Determination of Regional Precipitation Climates. *Journal of Climate*, 6, 2309-25.

Hosking, J. R. M. (1990). L-moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. *Journal of the Royal Statistical Society (Series B)*, Vol. 52, No. 1, pp. 105 – 124. ISSN 1467-9868.

Hosking, J. R. M., Wales, J. R. (1997). *Regional frequency analysis: An Approach Based on L-moments*. 1st ed. New York: Cambridge University Press, 209 p. ISBN 0-521-43045-3.

Kyselý, J., Picek J. (2007). Regional Growth Curves and Improved design Value Estimates of Extréme Precipitation Events in the Czech Republic. *Climate Research*, Vol. 33, pp. 243 – 255. ISSN 1616-1572.

Miskolczi, M. Langhamrová, J. (2011). Labor market and simultaneous equations solved by TSLS. Bratislava February 1 – 4, 2011, Proceedings. In: 10th International Conference *APLIMAT 2011*, Bratislava: Slovak University of Technology, Bratislava, pages 1597 – 1608. ISBN 978-80-89313-52-5.

Sipková, L`., Sodomová, E. (2009). Income distribution model in the Slovak Republic used the household SILC data. *Wspólczesne problemy modelowania i prognozowania zjawisk spoleczno-gospodarcznych*. Uniwersytet Ekonomicznego w Krakowie, pp. 241 – 256.

Contact

Diana Bílková

University of Economics in Prague

Faculty of Informatics and Statistics

Department of Statistics and Probability

bilkova@vse.cz