MODELLING OF INCOME AND WAGE DISTRIBUTION USING THE METHOD OF L-MOMENTS OF PARAMETER ESTIMATION

Diana Bílková

Abstract

Using L-moments is theoretically preferable to the conventional moments and consists in the fact that L-moments characterize a wider range of distribution. When estimating from sample L-moments, L-moments are more robust to the presence of outliers in the data. Experience also shows that, compared to conventional moments, L-moments are less prone to bias of estimation. Parameter estimates obtained using L-moments are mainly in the case of small samples often even more accurate than estimates of parameters made by maximum likelihood method. Using the method of L-moments in the case of small data sets from the meteorology is primarily known in statistical literature. This paper deals with the use of L-moments in the case for large data sets of income distribution (individual data) and wage distribution (data are ordered to form of interval frequency distribution of extreme open intervals). This paper also presents a comparison of the accuracy of the method of L-moments with an accuracy of other methods of point estimation of parameters of parametric probability distribution in the case of large data sets of individual data and data ordered to form of interval frequency distribution. Three-parametric lognormal curves were used as the model in all cases.

Key words: L-moments, sample L-moments, three-parametric lognormal distribution, methods of parameter estimation

JEL Code: C13, C16

Introduction

The applicability of the estimates of income and wage distribution is that it provides the possibility of linking the considerations relating to income and wage differentiation with socio-political considerations, in which it is not mostly enough to estimate development of the average income and wage, but it is necessary to estimate the proportions of workers

with low, middle and high incomes and wages or it is necessary to estimate the proportions of workers in all income or wage groups. Knowledge of models of income and wage distribution is also used for example in assessing the population\s living standards or at inter-area and international comparisons of living standards. In the field of statistics, we see many more using the knowledge of the income and wage distribution.

Commonly used statistical procedure to describe the observed statistics sets is to use their conventional moments or cumulants. Also, when choosing an appropriate parametric distribution for the data file, the parameters of the parametric distribution are usually estimated using the moment method of parameter estimation, which consists in creating equations in which sample conventional moments lay in the equality of the corresponding moments of the theoretical distribution. However, the moment method of parameter estimation is not always convenient, especially for small samples.

An alternative approach is based on the use of other characteristics, which we call L-moments, which are analogous to conventional moments, but they are based on linear combinations of order statistics, i.e. L-statistics. Using L-moments is theoretically preferable to the conventional moments, which consists in the fact that L-moments characterize a wider range of distribution. L-moments are more robust to the presence of outliers in the data when estimating from a sample. Experience also shows that L-moments are less prone to estimation bias compared with conventional moments and in finite samples, they are closer to asymptotical normal distribution. Parameter estimates obtained using the L-moment method are often even more accurate than parameter estimates made by maximum likelihood method, especially in the case of small samples.

From the statistical literature it is well-known 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 year net household income per capita (in CZK), and second, data sorted into a form of interval frequency distribution, these data refer to gross monthly wage (in CZK). 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 SILC and Microcensus 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 were compared with the accuracy of other

methods of parameter estimation, such as moment method, quantile method, maximum likelihood method.

1 L-Moments of Probability Distributions

Suppose that X is real random variable with distribution function $F(x)$ and with quantile function $x(F)$ and that $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$ are order statistics of random sample of sample size n, which is taken from the distribution of variable X. Then the r-th L-moment of random variable X is defined

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

Natural L-moment estimate λ_r based on the observed data sample is a linear combination of ordered data values, i.e. so called L-statistics. The expected value of order statistics has the form

$$EX_{j:r} = \frac{r!}{(j-1)! \cdot (r-j)!} \cdot \int_0^1 x [F(x)]^{j-1} \cdot [1-F(x)]^{r-j} \, \mathrm{d}F(x). \tag{2}$$

It is valid for the first four L-moments

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

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

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

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

The so called coefficients of L-moments are defined

$$\tau_r = \frac{\lambda_r}{\lambda_2}, \quad r = 3, 4, 5, ... \tag{7}$$

L-moments λ_1, λ_2, λ_3, …, λ_r and coefficients of L-moments τ_1, τ_2, τ_3, …, τ_r can be used as the characteristics of a distribution. In particular, L-moments λ_1 and λ_2 are considered as the characteristics of location and variability and coefficients of L-moments τ_3 a τ_4 are considered as the characteristics of skewness and kurtosis.

The three-parametric lognormal distribution LN(μ, σ^2, ξ) is described in detail for example in (Bartošová, 2006) or (Bílková, 2008). Using relations (3) to (5) and using equation (7) we obtain the first three L-moments of three-parametric lognormal distribution. It is valid for these L-moments

$$\lambda_1 = \xi + \exp\left(\mu + \frac{\sigma^2}{2}\right), \tag{8}$$

$$\lambda_2 = \exp\left(\mu + \frac{\sigma^2}{2}\right) \cdot \operatorname{erf}\left(\frac{\sigma}{2}\right), \tag{9}$$

$$\tau_3 = \frac{6\,\pi^{-1/2}}{\operatorname{erf}\left(\frac{\sigma}{2}\right)} \cdot \int_0^{\sigma/2} \operatorname{erf}\left(\frac{x}{\sqrt{3}}\right) \cdot \exp\left(-x^2\right) \mathrm{d}\,x, \tag{10}$$

2 Sample L-Moments

We 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. Then the r-th sample L-moment is defined

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

Especially it is valid for the firs four sample L-moments

$$l_1 = n^{-1} \cdot \sum_i x_{i:n}, \tag{12}$$

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

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

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

Sample L-moments can be used like as conventional sample moments, because they are characteristics of the basic properties of a sample distribution, i.e. location, variability, skewness and kurtosis, and they estimate the corresponding features of the probability distribution, from which were the data sampled. They can therefore be used to estimate the parameters of the basic probability distribution. In these cases, the L-moments are often preferred over conventional moments, because as a linear function of data they are less sensitive to sample variability and to size of errors in the case of outliers in the data than conventional moments. Therefore, we expect that they provide more accurate and robust estimates of the characteristics or parameters of the basic probability distributions. L-moments are described in detail for example in (Guttman, 1993), (Hosking, 1990), (Hosking, Wales, 1997) or (Kyselý, Picek, 2007).

3 Parameter Estimation

Let a distribution function of standardized normal distribution Φ, then Φ^{-1} is a quantile function of standardized normal distribution. It is valid for a 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 coefficient of L-moments (7) are usually estimated by

$$t_r = \frac{l_r}{l_2}, \quad r = 3, 4, 5, ... \tag{17}$$

Now we take the parameter estimates of three-parametric lognormal distribution as

$$z = \sqrt{\frac{8}{3}} \cdot \Phi^{-1}\left(\frac{1+t_3}{2}\right), \tag{18}$$

$$\hat{\sigma} \approx 0{,}999\,281 z - 0{,}006\,118 z^3 + 0{,}000\,127 z^5, \tag{19}$$

$$\hat{\mu} = \ln\left[\frac{l_2}{\operatorname{erf}\left(\frac{\sigma}{2}\right)}\right] - \frac{\hat{\sigma}^2}{2}, \tag{20}$$

$$\xi = l_1 - \exp\left(\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right), \tag{21}$$

4 Suitability of the Constructed 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.

5 Outputs

Tab. 1 contains calculated values of sample L-moments, the estimated parameters of the three-parametric lognormal distribution obtained using the L-moment method and the sum of absolute deviations of the observed and theoretical frequencies that the model assumes *S*. Tab. 1 refers to the distribution of the net year household income per capita. Tab. 2 presents the same for the distribution of gross monthly wage. For comparison, Tab. 3 contains the estimated parameters of the three-parametric lognormal distribution, which were acquired by moment method of parameter estimation and the sum of all deviations of the observed and theoretical frequencies for all intervals *S*, both for the distribution of the net year household income per capita and for the distribution of the gross monthly wage. The moment method of parameter estimation is described for example in Bílková, 2008). We can see form this table that the value of the parameter ξ (beginning of the distribution) can be negative. This means that the initially course of this curve gets into negative territory. This does not interfere with a good agreement of the model with the actual distribution due to the fact that the curve

is initially very close contact with the horizontal axis. Parameter ξ cannot give any interpretation for its negative values. It should be noted here that the purpose of this study is not to compare these two files with each other, but the purpose is to investigate the accuracy of parameter estimation for different types of data in terms of their arrangement within the

Tab. 1: Sample L-moments and estimated parameters of the lognormal distribution using the method of L-moments − distribution of net year household income per capita

Year	Sample L-moments			Estimated parameters			*S*
	l_1	l_2	l_3	μ	σ^2	ξ	
1992	35,246.51	7,874.26	2,622.14	9.696	0.490	14,491.687	1,904.506
1996	66,121.92	16,237.54	5,685.45	10.343	0.545	25,362.753	1,616.537
2002	105,029.89	27,978.40	10,229.62	10.819	0.598	37,685.637	625.662
2005	111,023.71	28,340.18	9,113.57	11.028	0.455	33,738.911	570.824
2006	114,945.08	28,800.68	9,286.18	11.040	0.458	36,606.903	1,336.021
2007	123,806.49	30,126.11	9,530.57	11.112	0.440	40,327.610	2,333.984
2008	132,877.19	31,078.96	9,702.45	11.163	0.428	45,634.578	2,639.240

Source: own research

Tab. 2: Sample L-moments and estimated parameters of the lognormal distribution using the method of L-moments − distribution of gross monthly wage

Year	Sample L-moments			Estimated parameters			*S*
	l_1	l_2	l_3	μ	σ^2	ξ	
2002	17,437.49	4,251.48	1,267.44	9.238	0.388	4,952.259	134,844
2003	18,663.18	4,524.95	1,251.90	9.402	0.332	4,364.869	135,841
2004	19,697.57	5,001.34	1,586.09	9.313	0.442	5,872.138	252,002
2005	20,738.14	5,262.93	1,636.67	9.392	0.424	5,908.390	260,423
2006	21,803.28	5,454.74	1,738.23	9.393	0.447	6,795.207	277,559
2007	23,882.83	6,577.65	2,627.93	9.222	0.724	9,349.280	429,282
2008	25,477.59	6,993.72	2,737.94	9.319	0.693	9,719.297	455,574

Source: own research

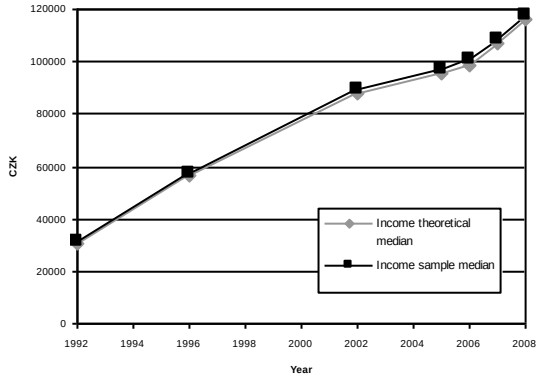
Tab. 3: Estimated parameters of the lognormal distribution using the moment method − distribution of net year household income per capita and distribution of gross monthly wage

Year	Income				Year	Wage			
	μ	σ^2	ξ	*S*		μ	σ^2	ξ	*S*
1992	8.883	1.083	22,284.335	2,985	**2002**	9.492	0.264	2,311.688	114,691
1996	9.154	1.334	45,269.967	4,161	**2003**	9.698	0.155	2,993.514	157,301

2002	9.668	1.327	66,925.879	2,418	**2004**	9.779	0.221	-25.695	226,646
2005	9.710	1.287	73,299.950	1,478	**2005**	9.906	0.193	-1,339.601	225,479
2006	9.976	1.177	71,936.249	2,281	**2006**	9.979	0.180	-1,805.527	248,955
2007	10.242	1.079	73,575.417	2,736	**2007**	9.734	0.377	3,509.924	332,148
2008	10.328	1.044	80,180.795	2,848	**2008**	9,851	0,345	2,920.381	341,796

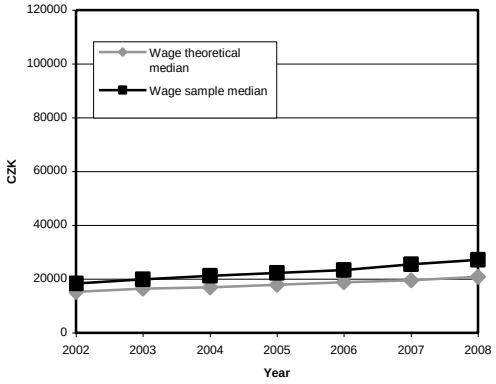
Source: own research

Fig. 1: Development of theoretical and sample median of the net income per capita



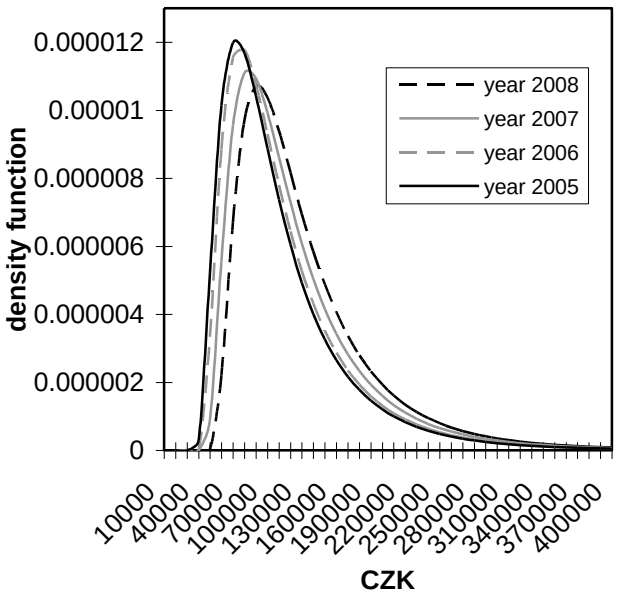
Source: own research

Fig. 2: Development of theoretical and sample median of the gross monthly wage



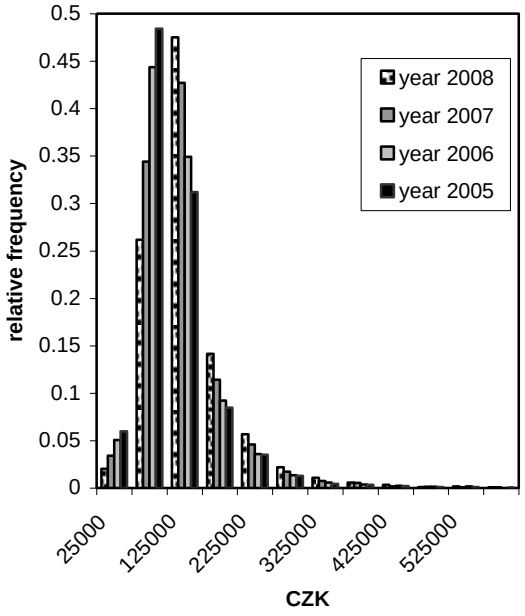
Source: own research

Fig. 3: Probability density function of the net income per capita (years 2005-2008)



Source: own research

Fig. 4: Frequency histogram of the net income per capita (years 2005-2008)



Source: own research

meaning of individual data and data organized to form of frequency distribution. Another purpose of this study is to compare the accuracy of different methods of parameter estimation with the accuracy of the L-moment method.

Fig. 1 represents the development of the sample and theoretical median of the three-parametric lognormal distribution with parameters estimated using the L-moment method for the distribution of the net year household income per capita and Fig. 2 represents the same for the distribution of gross monthly wage. Fig. 3 contains the development of probability density function (in the years 2005-2008) of the theoretical three-parametric lognormal distribution with the parameters estimated using the L-moment method for the distribution of the net household income per capita and Fig. 4 presents the corresponding sample interval frequency distribution.

The values of well known test criterion χ^2 were also calculated, but due to the fact that in these large sample sizes, such as in the case of income and wage distribution are seen, the test power is too high that test uncovers the all very slight deviations between the sample and theoretical distribution. This test results to the rejection of the tested hypothesis about the expected theoretical distribution practically in all cases. However, we are not interested in such small deviations and approximate agreement between model and reality is sufficient. For this reason, we do not give the values of the test criterion χ^2.

We can see from Tabs. 1 − 3 that the values of *S* are considerably higher in the case of data set arranged to the form of interval frequency distribution (distribution of gross monthly wage) than in the case of individual data set (distribution of net year household income per capita), which was expected. We can also see that the values *S* result essentially higher in the case of moment method of parameter estimation than in the case of L-moment method − both regarding to the set of individual data. But we cannot say the same thing in terms of data into a form of interval frequency distribution, where the value *S* results comparable in the case of both data sets. If we compare the accuracy of the method of L-moments with an accuracy of other methods of parameter estimation (quantile method and even the maximum likelihood method), we come to similar conclusions as to the accuracy of this method compared with the accuracy of moment method.

6 Conclusions

The L-moment method of parameter estimation gives more accurate results than other methods of parameter estimation (moment method, moment method, maximum likelihood

method) for individual data. In the case of data grouped to form of interval frequency distribution, all four methods of parameter estimation offer comparable results. In these cases, the inaccuracies arise above all at both tails of the distribution (heavy tails). All Figs. 1 − 4 are related to the L-moment method of parameter estimation and they also give an idea about the accuracy of this method.

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

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.

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.

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.

Contact

Diana Bílková

University of Economics in Prague

Faculty of Informatics and Statistics

Department of Statistics and Probability

nám. W. Churchilla 4, Praha, Czech Republic

bilkova@vse.cz