AUSTRIAN JOURNAL OF STATISTICS
Volume 41 (2012), Number 2, 125–132

Application of the L-Moment Method when Modelling the Income Distribution in the Czech Republic

Diana Bílková and Ivana Malá
University of Economics, Prague

Abstract: This paper deals with modelling income distributions in the Czech Republic in 1992–2007. The net annual income per capita for Czech households is evaluated from data based on the microcensus and the EU-SILC 2005–2008. For all analysed years the distribution of incomes was estimated in the whole sample as well as in the subgroup of households, whose heads are physicists (or experts in related sciences), architects and engineers. In the paper the three-parametric lognormal distribution is used as a model. Unknown parameters are estimated with the use of four methods – those of maximum likelihood, quantiles, moments and L-moments.

Zusammenfassung: Der Artikel befasst sich mit der Modellierung der Einkommen in der Tschechischen Republik in den Jahren 1992–2007. Das Nettojahreseinkommen pro Person für die tschechischen Haushalte ist aus Erhebungen des Mikrocensus und des EU-SILC 2005–2008 ausgewertet. Für alle untersuchten Jahre wird die Verteilung der Einkommen in der gesamten Stichprobe als auch in Untergruppen jener Haushalte analysiert, deren Vorstände wissenschaftliche Mitarbeiter oder Fachmänner auf dem Gebiet der physikalischen oder verwandten Wissenschaften, Architekten und Ingenieure stehen. Als Modell für die Verteilung wird die dreiparametrige Lognormalverteilung verwendet. Die Schätzung der unbekannten Parametern erfolgt durch vier Verfahren: durch die Maximum-Likelihood Methode, Quantilsmethode, Momentenmethode und durch die L-Momentmethode.

Keywords: Tantile, Lognormal Distribution, L-Moment, Maximum Likelihood Estimation.

1 Introduction

Statistical procedures commonly used for the description of the observed statistical sets lie in the use of their conventional moments, cumulants or quantiles. An alternative approach is based on the use of other moment characteristics called L-moments. They are analogous to conventional moments but based on linear combinations of order statistics. The use of L-moments is appropriate from both theoretical and practical points of view. L-moments characterize a wider range of distributions than classical moments since a finite expected value implies all L-moments as finite. Moreover, 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 an asymptotic normal distribution. Parameter estimates obtained with the use of L-moment method are often even more accurate than parameter estimates made by the maximum likelihood method (this method is theoretically optimal for large sample sizes), especially in the case of small samples; see Hosking (1990).

The knowledge of the probability distribution (or at least its model) gives more detailed information about incomes than the characteristics of location, variability or shape. In the text, we fit the three-parametric lognormal distribution both to the whole sample and the subgroup of interest. The appropriateness of using the theoretical distribution for this purpose is explained in the statistical literature; see e.g. Bílková (2008) or Kleiber and Kotz (2003).

The issue of income distribution is treated extensively, distributions and methods with applications to national data being widely discussed in Kleiber and Kotz (2003). Incomes in the Czech Republic and Slovakia are analysed, for example, in Bartošová (2009) or Bartošová and Forbelská (2011)). All methods of estimation that are used in this paper, including the three-parametric lognormal distribution, are described in the statistical literature; see e.g. Bílková (2008). The three-parameter lognormal distribution is discussed in detail, for example, in Bartošová and Bína (2009) or in Bílková (2008), a moment method of parameter estimation in Bartošová (2009) or Bílková (2008), quantile method in Bílková (2008) or Sipková and Sodomová (2009), maximum likelihood method in Bílková and Malá (2010). The concept of L-moments and the use of these quantities in the estimation of parameters of probability distributions can be found in Bílková (2011), Hosking (1990) or Hosking and Wales (1997).

This paper deals with modelling of distributions of net annual household per capita (nominal) income of Czech households. In the analysis, the lognormal distribution is used and the results of four different parameter estimation methods (those of moments, quantile, maximum likelihood and L-moments) are compared. Data from Microcensus (1992, 1996, 2002) and the EU Statistics on Income and Living Conditions (EU-SILC 2005–2008) surveys conducted by the Czech Statistical Office are used in the text, covering the period of 16 years (1992–2008).

The lognormal distribution is fitted into the sample and a subgroup of households including those whose heads are creative workers (“scientists and experts in physics and related sciences, architects and engineers”).

2 Data and Results

In the paper, data from Microcensus (1992, 1996, 2002) and the European Union Statistics on Income and Living Conditions (EU-SILC 2005–2008) surveys carried out by the Czech Statistical Office are used. The surveys held in 2005–2008 cover incomes from 2004 to 2007. The first sample (from 1992) represents the part of the survey dealing with households in the Czech and Slovak Federal Republic, the country that split into the Czech Republic and the Slovak Republic in 1993.

In all samples, a total annual net income per capita for each household is evaluated as the total net income of a household divided by the number of its members. From the various characteristics of households, only head of household’s occupation is used. In addition to the estimates in the whole population of Czech households, special attention was paid to a subgroup of households whose heads are classified as “creative workers”. This subgroup consists of scientists and experts in physics, chemistry, mathematics, statistics and informatics, designers, architects, constructors and other related branches (CZSO, Czech statistical office). Members of this group have acquired technical university degrees, so a strong positive impact on their income and, consequently, their households' income seems to be predictable. Focusing exclusively on the head of household's job, other household members' (especially spouses') field of occupation is not taken into account. The defined subgroup of households is relatively small in comparison to the whole sample (Table 2).

Mean and median, together with less frequently used medial, are used as the characteristics of location. The medial is the value of a 50 % (sample) tantile just as the sample median equals the value of a 50 % sample quantile. Sample tantiles as well as sample quantiles are based on an ordered sample. First of all, cumulative sums of observations in the ordered sample are evaluated. Then, for a given percentage p, $0 < p < 100$, a p % tantile is defined as the value of the analysed variable that divides all observations in the ordered sample into two parts: the sum of smaller or equal observations is p % of the total sum of observations and the sum of observations that are greater represents the residual $(100 - p)$ percent of this sum. It can be derived from this definition that the medial can be used as a reasonable characteristic of the level of income, since households with the income lower or equal to the medial receive one half of the total income in the sample, those with the income higher than the medial receiving the other half.

Table 1 presents the values of sample characteristics of location, variability and shape of the distribution of net annual household income per capita (in CZK) only for the last analysed year 2007. The difference in the location between the total sample and the subsample is well notable. We can see remarkably higher values of all three characteristics of location (mean, median, medial) in the analysed subgroup than in the total sample of households in 2007. Moreover, we observe that the medial is the highest of these characteristics, followed by the average, the median being the lowest. The sample standard deviation is markedly higher in the analysed subgroup (it implies a bigger difference in incomes), but the coefficient of variation, as the characteristic of relative variability, is slightly lower in the subgroup. The frequency distribution of the entire sample of households has substantially greater skewness and kurtosis than the frequency income distribution of the analysed subgroup. Moreover, we can see selected quantiles in Table 1. Lower and upper quartiles define the interval of middle 50 % of incomes. The first decile means the upper limit of 10 % of the lowest incomes, the ninth decile describing the lower limit of 10 % of the highest incomes. One half of the length of the interval between quartiles (quartile deviation) can be used as a quantile characteristic of variability. Its value is 25,889 CZK for the whole sample and more than twice higher (58,889 CZK) for the analysed subgroup. We can see comparatively low quantiles (a difference of 19 % for the first decile and 13 % for the first quartile), the difference being higher in upper quantiles (54 % for the upper quartile and 64 % for the last decile). It indicates differences in right tails of distributions. Sample characteristics of location are shown (together with other values) for all analysed years in Figure 1.

Table 2 contains cumulative inflation rates between analysed years; values for the period 1992 – 1996 are not available on the web pages of the Czech Statistical Office. Table 2 additionally presents sample sizes and values of the first three sample L-moments for

Table 1: Sample characteristics (in CZK) of income per capita in 2007.

Sample Characteristic	Total Set	Analysed Subgroup
Arithmetic Mean	132,877	187,615
Median	117,497	155,712
Medial	133,930	212,733
Standard Deviation	73,982	102,601
Coefficient of Variation	0.5568	0.5469
Skewness	6.979	1.375
Kurtosis	123.826	1.750
1^{st} decile	76,571	91,105
Lower quartile	97,160	109,986
Upper quartile	148,937	229,764
9^{th} decile	202,327	331,759

Source: own computations

Table 2: Inflation rate, sample sizes (n), first three sample L-moments (l_1, l_2, l_3) in 1992–2007 (CZK).

		Total Set				Analysed Subgroup			
Year	Inflation	n	l_1	l_2	l_3	n	l_1	l_2	l_3
1992	—	16,233	35,246	7,874	2,622	231	55,752	17,788	7,766
1996	—	28,148	66,121	16,237	5,682	320	66,121	16,237	5,685
2002	1.358	7,973	105,029	27,978	10,229	130	149,918	45,569	19,191
2004	1.029	4,351	111,023	28,340	9,113	81	146,872	37,982	4,621
2005	1.019	7,483	114,945	28,800	9,286	107	174,383	50,038	13,974
2006	1.025	9,675	123,806	30,126	9,530	148	182,063	54,067	13,212
2007	1.028	11,294	132,877	31,078	9,702	169	187,615	54,076	15,271

Source: own computations

the whole sample (left part) and the analysed subgroup (right part) respectively. The first sample L-moment coincides with the average of values (theoretical first L-moment with an expected value), the second is a characteristic of variability, the third marks skewness of empirical distributions. The development of characteristics during the analysed period is obvious from the table. In 2004 (first survey held according to the EU methodology), it is rather different (also giving strange value for shift parameter of the lognormal distribution), but there are only 81 observations in the subgroup.

Table 3 presents the estimated values of the parameters of the three-parameter lognormal distribution. All parameters have a simple probabilistic interpretation. If X means a random variable with the three-parameter lognormal distribution, then the parameter μ denotes the expected value of the random variable $\log(X - \theta)$, σ^2 is the variance of this variable and the shift parameter θ describes a theoretical minimal value of the distribution in the form $X > \theta$. We can see from the table that the value of the parameter θ (theoretical minimum) is, in many cases, negative. This, however, does not cause a problem for

Table 3: Estimated parameters of the three-parameter lognormal distribution in 2007.

		Parameter Estimates		
Method of	Set	μ	σ	θ
Moments	Total set	10.328	1.044	80,179
	Analysed subgroup	12.293	0.414	$-50,073$
Quantiles	Total set	10.961	0.646	59,909
	Analysed subgroup	11.691	0.715	36,176
Maximum likelihood	Total set	11.703	0.421	−171
	Analysed subgroup	11.855	0.767	20,145
L-moment	Total set	11.163	0.654	45,635
	Analysed subgroup	12.010	0.711	146

Source: own computations

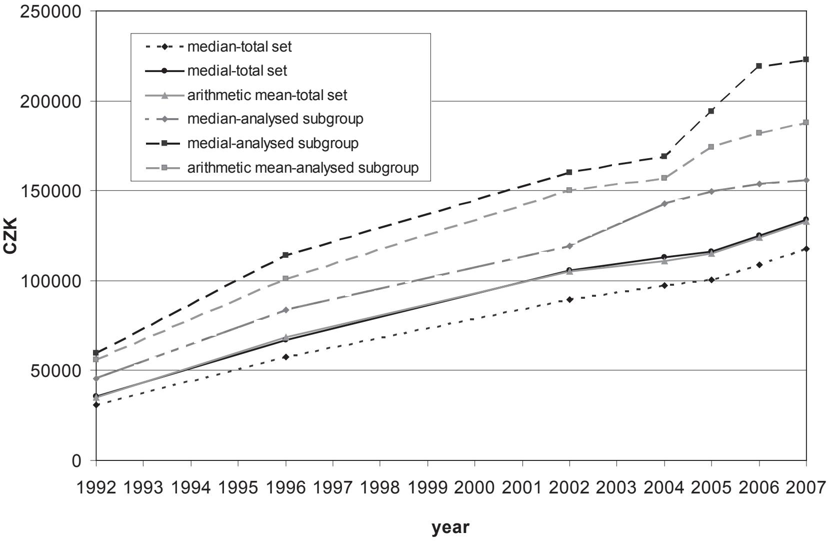

Figure 1: Characteristics of location (in CZK).

a good fit of the model, since the beginning of the three-parametric lognormal curve lies very close to the horizontal axis. Nevertheless, the above mentioned negative values are not interpreted.

The development of location characteristics from 1992 to 2007 is shown in Figure 1 that displays characteristics of nominal incomes, the inflation rate being given in Table 2. We can see in the figure that all characteristics of location in the analysed subgroup lie substantially higher than those of the whole set of households throughout the analysed period. For the positively skewed distribution the following order of characteristics of location is valid: median is the smallest, then the average follows and medial has the highest value. It is observable from the figure, too (see also Table 1).

The estimated three-parameter lognormal probability densities are shown in Figure 2. There are surprisingly large differences between the probability densities estimated with the use of the four methods of estimation. The maximum likelihood method is – in theory

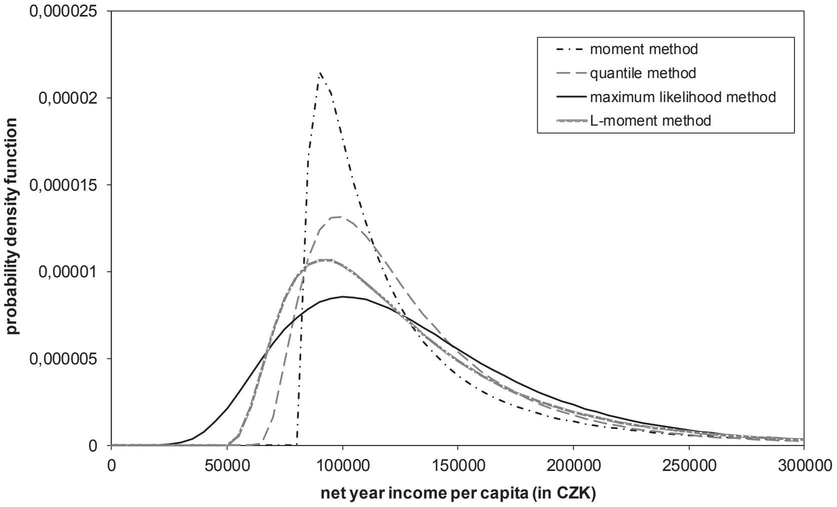

Figure 2: Estimated probability densities of the three-parameter lognormal distribution for the whole population in 2007.

– asymptotically optimal: the estimators are consistent, efficient and (asymptotically) unbiased. Comparable results should be given by the L-moments method which might be better for small subgroup samples. Having employed these two methods, we can see close values of estimates, – with the exception of the shift parameter theta – from Table 3. Both methods of moments and quantiles are easily applicable but not precise enough, thus giving reliable results only for large samples, as such estimators are consistent.

The question of suitability of the chosen lognormal distribution is not a common statistical problem of testing the null hypothesis "H_0: The sample comes from the assumed distribution" versus the alternative hypothesis: "H_1: not H_0". In the case of a goodness of fit test for income distributions, we often encounter the large n problem. When working with large data sets, the test tends to reject almost all null hypotheses. There are two reasons for this. First, the power of this test is too high for large samples (for a given significance level), taking into account even the smallest differences in the real income distribution and the model. The other reason is the principle of test construction itself. Since the small differences are beyond our scope, an approximate curve fit is sufficient – we "just borrow the model (the curve)". In these cases, the use of the well-known χ^2 criterion is rather limited. Therefore, the interpretation of the results can be more or less arbitrary and we have to embrace experience and logical analysis.

Figure 3 indicates how to compare the accuracy of parameter estimation methods. It includes the developments of sample median and theoretical medians of the three-parameter lognormal distribution (evaluated from the estimated distributions, having used various methods for the period 1992–2007). There are only three methods depicted in the figure. The quantile method is an exception, because the median is chosen as one of the three quantiles for the estimation and the estimated medians would coincide with the sample values. In the figure, the medians evaluated with the use of L-moment estimates are closest to empirical medians, those from moment method being most distant from the empirical values.

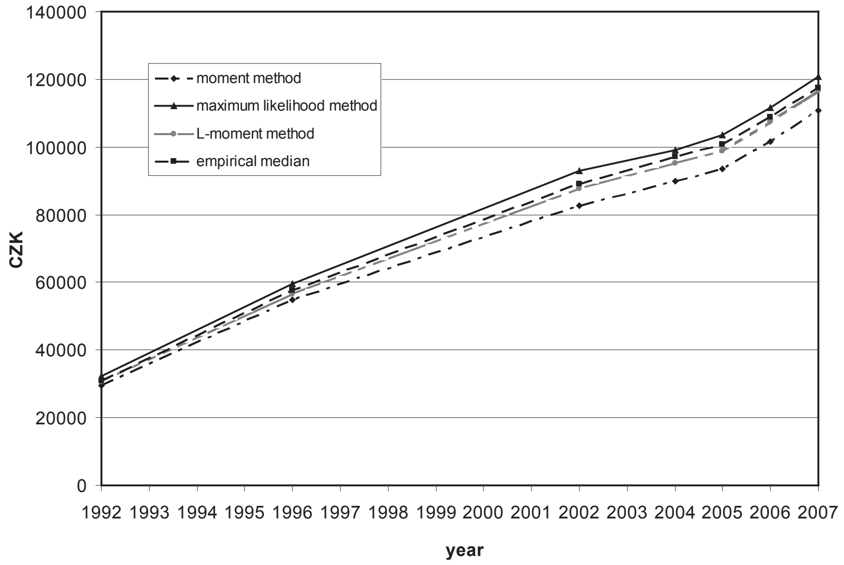

Figure 3: Sample and theoretical median of the income per capita in 1992–2007.

3 Conclusion

The paper presents an analysis of incomes in the Czech Republic in 1992–2007, using the lognormal distribution.

In order to show excellent properties of estimation with the use of L-moments, we applied three frequently used methods of parameter estimation – maximum likelihood, moment and quantile methods. Maximum likelihood is theoretically optimal for large samples, the other methods being simple but not very precise (especially in the case of small samples); their results are sometimes used as an initial approximation for numeric search for maximum likelihood estimates. These theoretical properties of methods of estimation are visible in our results. The results from L-moments estimation proved to be highly satisfactory. They are consistent with the results presented in the statistical literature (for small samples in particular); see e.g. Hosking (1990) proving that the L-moment method gives the most accurate results.

The results obtained by employing various methods of parameter estimation differ considerably even for large samples (Figure 2). The moment method brings a really different model, the results from other methods being comparable. Moreover, characteristics evaluated from these estimated distributions are very similar (see e.g. Figure 3).

In the analysed subgroup, there are only highly qualified household heads with university degrees. The results of our research show that the level of income in these households headed by professionals is remarkably higher than that in the total population of households. The positive impact of technical university education and creative work on the level of incomes has been quantified.

The absolute variability is markedly greater for the analysed subgroup, which is probably caused by more job opportunities for highly qualified people, the relative variability (coefficient of variation) being comparable with the total set.

Acknowledgements

The authors were supported by grant project IGS 24/2010 from the Faculty of Informatics of the University of Economics in Prague.

References

Bartošová, J. (2009). Analysis and modelling of financial power of Czech households. In *8th International Conference APLIMAT 2009, Bratislava* (p. 717-722). Bratislava: Slovak University of Technology.

Bartošová, J., and Bína, V. (2009). Modelling of income distribution of Czech households in years 1996–2005. *Acta Oeconomica Pragensia*, *17*, 3-18.

Bartošová, J., and Forbelská, M. (2011). Differentiation and dynamics of household incomes in the Czech EU-SILC survey in the years 2005–2008. In *10th International Conference APLIMAT 2011, Bratislava* (p. 1451-1460). Bratislava: Slovak University of Technology.

Bílková, D. (2008). Application of lognormal curves in modeling of wage distributions. In *7th International Conference APLIMAT 2008* (p. 341-351). Bratislava: Slovak University of Technology.

Bílková, D. (2011). Use of the L-moments method in modeling the wage distribution. In *10th International Conference APLIMAT 2011, Bratislava* (p. 1471-1481). Bratislava: Slovak University of Technology.

Bílková, D., and Malá, I. (2010). Development of income distributions of households in the Czech republic since 1992. In *Prague Stochastics 2010.* Prague: Faculty of Mathematics and Physics of Charles University in Prague. (Poster)

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*, *52*, 105-124.

Hosking, J. R. M., and Wales, J. R. (1997). *Regional Frequency Analysis: An Approach Based on L-Moments*. New York: Cambridge University Press.

Kleiber, C., and Kotz, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. New York: Wiley-Interscience.

Sipková, L., and Sodomová, E. (2009). *Income distribution model in the Slovak republic used the household SILC data.* Uniwersytet Ekonomicznego w Krakowie.

Authors' address:

Diana Bílková and Ivana Malá

Department of Statistics and Probability
Faculty of Informatics and Statistics
University of Economics in Prague
Sq. W. Churchill 1938/4
130 67 Prague, Czech Republic

E-Mails: bilkova@vse.cz, malai@vse.cz
Web Page: http://www.vse.cz/