MODELLING INCOME DISTRIBUTION IN THE CZECH REPUBLIC WITH A SPECIAL FOCUS ON INCOME DISTRIBUTION OF PROFESSIONALS IN INFORMATICS

Diana Bílková – Ivana Malá

Abstract

This paper deals with the modeling of income distribution in the Czech Republic in 2008. The values of the net year household income per capita (in CZK) are evaluated from data based on the Results of the Living Conditions Survey (a national module of the European Union Statistics on Income and Living Conditions (EU-SILC)) that has been organized yearly by the Czech Statistical Office since 2005. Data for the whole sample are analyzed and the comparison with the group of professionals in informatics is given. In the survey Czech households are taken as statistical units and we will analyze an occupation of a head of a household. The group of occupations in informatics is not included in the questionnaire the closest groups are scientists and experts in physics and related sciences, architects and engineers (the creative workers)". This paper focuses on the use of lognormal distribution. In this paper moment and quantile methods, maximum likelihood or L-moment methods are used to estimate parameters of lognormal distribution. Using various methods we obtain different estimates of parameters of the theoretical distribution. The magnitude of these differences is shown in the paper together with the impact of these differences on characteristics of income.

Key words: Median, medial, tantiles, lognormal distribution, L-moment method

JEL Code: C13, C16

Introduction

This paper deals with the modeling of income distributions of the Czech households. In the analysis the lognormal distribution is used and results of four different methods of parameter estimation (method of moments, quantile method, maximum likelihood method, and not least method of L-moments) are compared. In the paper data from surveys Microcensus (1992, 1996, 2002) and European Union Statistics on Income and Living Conditions (EU-SILC) organised by the Czech Statistical Office are used. In these surveys a households are sampled, from such a data annual total net income per capita for each household is evaluated as the total income of a household divided by the total of members. In addition to the estimates in the whole population of Czech households the special attention was paid to a group of households with the head´s occupation classification characterise as "scientists and experts in physics and related sciences, architects and engineers (the creative workers)". This group includes people with the education in informatics, which are not researched separately.

For the description of samples usual characteristics of location, variability and shape of distributions are used in (Bílková, Budinský, Vohánka, 2009). We add less frequently used characteristic of location – medial. By Medial is a value of the 50% tantile. Sample tantiles, as well as sample quantiles, are based on the ordered sample. First of all cumulative sums of observations in the ordered sample are evaluated. Then for given percentage P ($0 < P < 100$) the P% tantile is defined as a value an analyzed variable that divide all observations in the ordered sample into two parts: the sum of smaller or equal observations is P % of the total sum of values and the sum of observations that are equal or greater represents the residual ($100 – P$) percent of this sum. It can be derive from this definition that the medial can be used as a reasonable characteristic of the level of income, because households with the income per capita lower or equal to the medial receive a half of the total income in the sample and households with the income greater or equal to the medial receive the other half of the total income.

The method of estimation parameters, including the three-parametric lognormal distribution are described in the statistical literature, see for example (Pacáková, Sipková, Sodomová, 2005) or (Sipková, Sodomová, 2009). Three-parametric lognormal distribution is discussed in detail for example in (Bartošová, Bína, 2009) or (Bílková, 2008), moment method of parameter estimation in (Bartošová, 2009) or (Bílková, 2008), quantile method in (Bartošová, Forbelská, 2011) or (Sipková, Sodomová, 2007), maximum likelihood method in (Bílková, Malá, 2010). The concept of L-moments and the use of these quantities in the estimation of parameters of probability distribution are given in (Bílková, 2011), (Hosking, 1990) or (Hosking, Wales, 1997).

The description of various models of income distribution and the discussion taht deals with the application of fitted income distributions in the evaluation of living standards is included in papers (Arltová, Langhamrová, 2009), (Arltová, Langhamrová, Langhamrová, 2010) or (Miskolczi, Langhamrová, 2011).

1. Results and Conclusions

Tab. 1 presents the values of sample characteristics of location, variability and shape of a distribution of net annual household income per capita (in CZK). The differences in analyzed samples are well notable. We can see significantly higher values of all three characteristics of location (mean, median, medial) in the group of scientists, and experts in physics and related sciences, architects and engineers than in the total sample of households in 2008. The sample standard deviation is significantly higher in the analysed subgroup (it implies bigger differences in incomes), but the coefficient of variation as the characteristic of relative variability is on the contrary slightly less in the subgroup. Frequency distribution of the total sample of households has substantially greater skewness and kurtosis than frequency income distribution of scientists and experts in physics and related sciences, architects and engineers. Moreover we can see selected quantiles in the Tab. 1. The lower and upper quartiles define an interval of middle 50 percents of incomes. The first percentile, is sometimes used as a characteristic of the minimum income and the ninety-ninth percentile can be used as a characteristic of the maximum income. The first decile means an upper limit of 10 % of the lowest incomes and the ninth decile describes the lower limit of 10 % of the highest incomes. We can see comparable low quantiles (first percentile, first decile, first quartile) and the difference is obvious in the upper quantiles as upper quartile, ninth decile or especially 99^{th} percentile. It shows people with very high incomes in the subgroup.

Tab. 1: Sample characteristics of income per capita (CZK) in 2008

Sample characteristic	Total set	Analysed subgroup
Arithmetic mean (in CZK)	132,877	187,615
Median (in CZK)	117,497	155,712
Medial (in CZK)	133,930	212,733
Standard deviation (in CZK)	73,982	102,601
Coefficient of variation (in %)	55.68	54.69
Moment skewness (without unit)	6.979	1.375
Moment kurtosis (without unit)	123.826	1.750
Lower quartile (in CZK)	97,160	109,986
Upper quartile (in CZK)	148,937	229,764

1st percentile (in CZK)	40,795	50,481
1st decile (in CZK)	76,571	91,105
9th decile (in CZK)	202,327	331,759
99th percentile (in CZK)	391,094	525,146

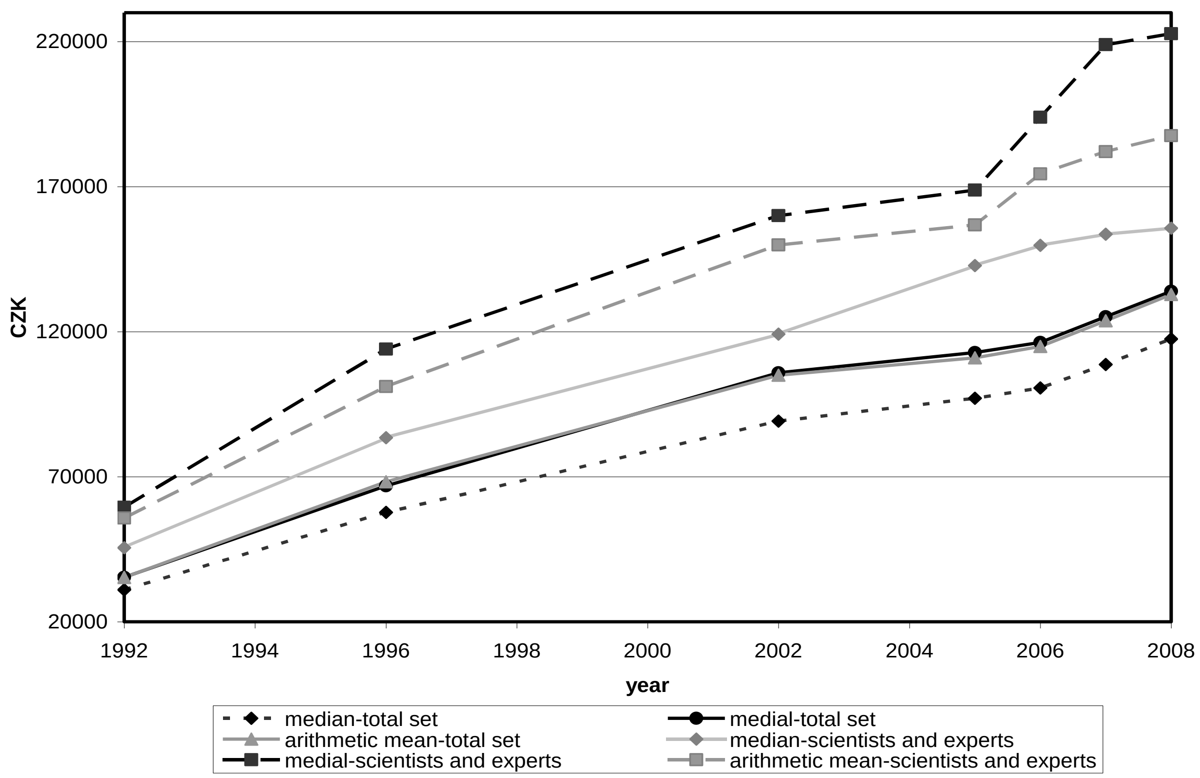
Source: own research

Tab. 2: First three sample L-moments (l_1, l_2, l_3) in 1992 − 2008

	Total set			Scientists and experts in physics and related sciences, architects and engineers		
Year	l_1	l_2	l_3	l_1	l_2	l_3
1992	89,766.165	28,487.340	6,662.317	55,752.596	17,788.276	7,766.527
1996	160,065.841	53,010.501	13,963.572	92,376.416	25,176.374	8,197.479
2002	105,029.886	27,978.395	10,229.621	149,917.969	45,569.737	19,191.463
2005	111,023.709	28,340.178	9,113.571	146,872.259	37,982.969	4,621.693
2006	114,945.080	28,800.682	9,286.184	174,383.606	50,038.874	13,974.047
2007	123,806.489	30,126.111	9,530.570	182,063.629	54,067.548	13,212.770
2008	132,877.187	31,078.956	9,702.450	187,615.074	54,076.639	15,271.027

Source: own research

Fig. 1: Development of the location characteristics (in CZK) in 1992 − 2008



Source: own research

Tab. 2 presents values of the first three sample L-moments for both populations. The first sample L-moment coincides with the mean value. The development during 17 analysed years is obvious in the table. In the year 2005 (first survey hold according to the EU methodology) is rather different (it gives also strange value for shift parameter of the lognormal distribution) but there are only 81 observations in the subgroup of analysed professionals.

Tab. 3: Estimated parameters of three-parametric lognormal distribution in 2008

Method	Set	Parameter estimations		
		μ	σ^2	θ
Moment method	**Total set**	10.328	1.089	80,179.043
	Analysed subgroup	12.293	0.171	−50,073.097
Quantile method	**Total set**	10.961	0.417	59,909.004
	Analysed subgroup	11.691	0.511	36,175.807
Maximum likelihood method	**Total set**	11.703	0.177	−171.167
	Analysed subgroup	11.855	0.589	20,144.659
L-moment method	**Total set**	11.163	0.428	45,634.578
	Analysed subgroup	12.010	0.506	146.052

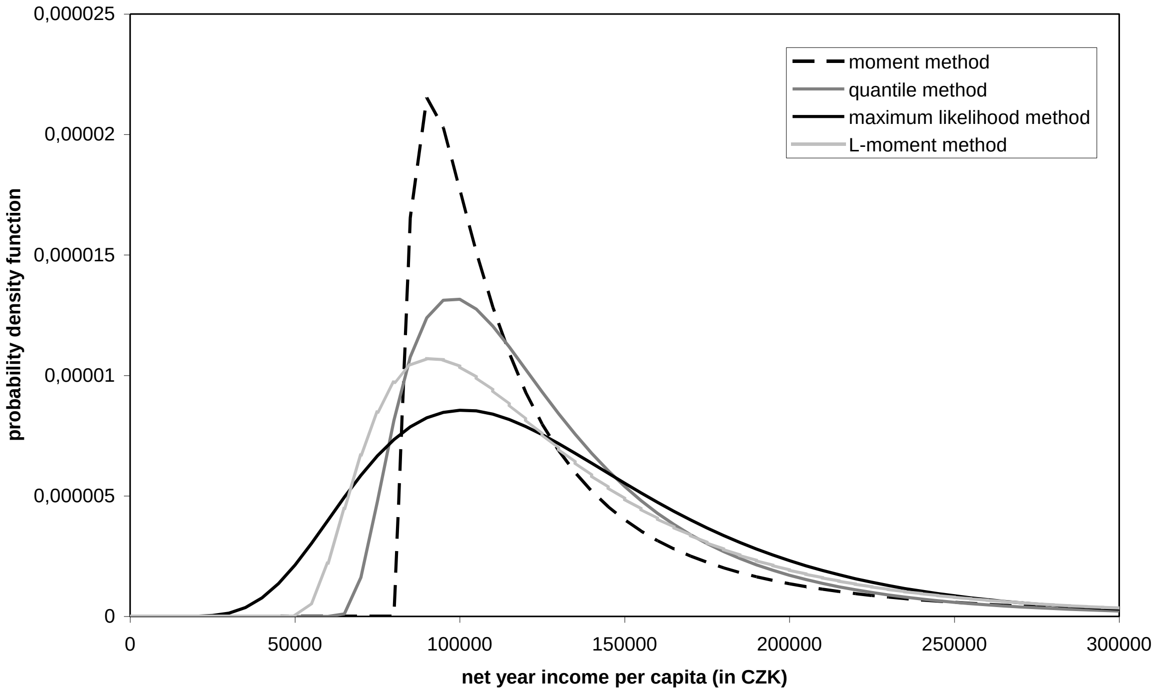
Source: own research

The development of the location characteristics from 1992 to 2008 is shown in Fig. 1. We can see in the figure that all the location income characteristics of the households of scientists and experts in physics and related sciences, architects and engineers lie substantially higher than the location income characteristics of the whole set of households throughout the monitored period.

Tab. 3 presents the estimated values of the parameters of three-parametric lognormal distribution. If X means a random variable with this distribution, than the parameter μ is an expectation of $\ln(X-\theta)$, σ^2 is a variance of this shifted distribution and a shift parameter θ

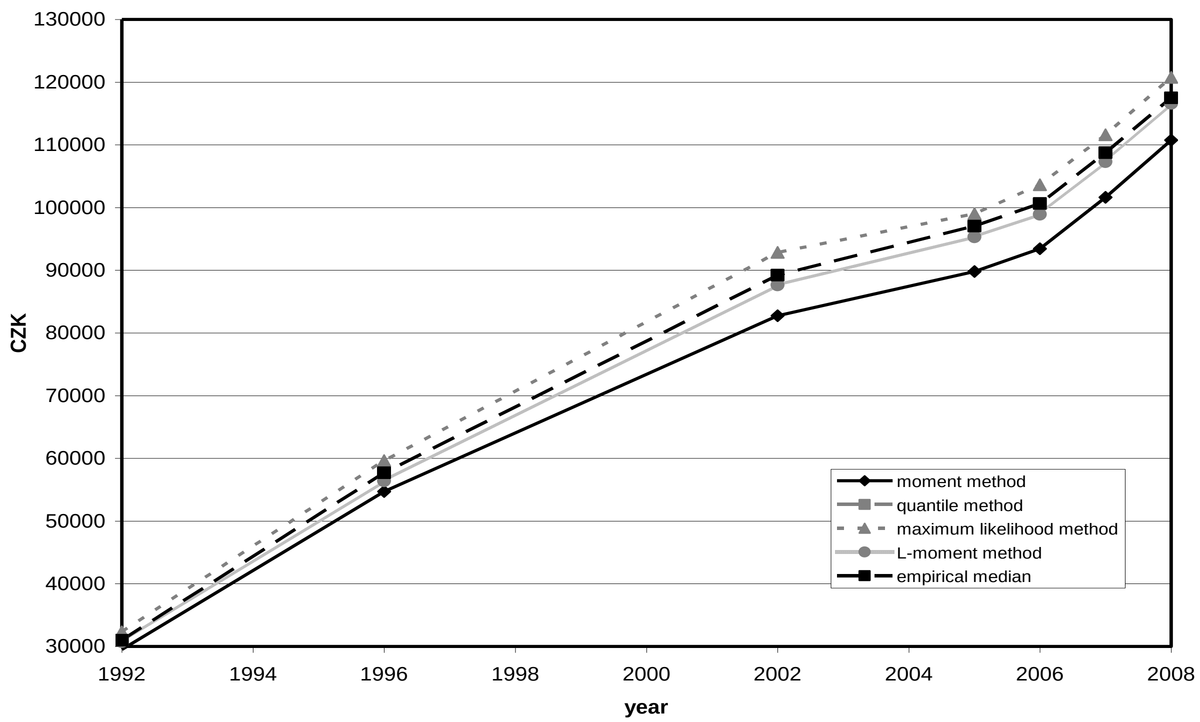
describes the minimal value of the distribution in the form $X > \theta$. We can see from this table that the value of parameter θ (theoretical minimum) is, in many cases, negative. However, since the three-parametric lognormal curve has very tight contact with horizontal axis at the beginning of its course, this fact does not have to be a problem for a good fit of the model.

Fig. 2: Estimated probability density functions of the three-parametric lognormal distribution for the whole population according to the method used in 2008



Source: own research

Fig. 3: Sample and theoretical median of the net annual income per capita in 1992 − 2008



Source: own research

The estimated three-parametric lognormal probability densities are shown in the Fig. 2. There are surprisingly obvious significant differences between the curves of probability density functions estimated with the use of various methods of estimation.

The question of suitability of a chosen log-normal 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: non H_0". In the case of the goodness of fit test for income distributions, we often encounter the large *n* problem. If we work with large datasets, the test tends to reject almost every null hypothesis. There are two reasons for this. Firstly the power of this test is for large samples (for a given significance level) too high and it takes into account even the smallest differences of the real income distribution and the model. The other reason is the principle of the test construction itself. Since small differences are out of our scope, the approximate fit of the curve 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 take the advantage of the experience and logical analysis.

The Fig. 3 provides some idea of the terms of accuracy of the methods of parameter estimation. It includes the development of sample median and the development of theoretical medians of three-parametric lognormal distribution (evaluated from the estimated distributions with the use of various methods for 1992 to 2008). The exception is the quantile method of parameter estimation, because we choose one of three quantile equations to be the equality of sample and theoretical median. Then theoretical median is always equal to the sample one and the development of a theoretical median of the quantile method coincides with the development of sample median and it tells us nothing about the relevancy of the lognormal curve. So with the exception of quantile method of parameter estimation, we can state that L-moment method of parameter estimation gives most accurate results, including maximum likelihood method. As expected, the moment method of parameter estimation gives the least accurate results.

Acknowledgment

The article was supported by grant IG 24/2010 from the University of Economics in Prague.

References

Arltová, M., Langhamrová, J. (2009). Analýza vztahů časových řad porodnosti a sňatečnosti v České republice v letech 1960-2007. Politická ekonomie, Vol. 4, No. 4, pp. 495–507. ISSN 0032-3233.

Arltová, M., Langhamrová, J. Langhamrová, J. (2010). Empirical Analysis of the Climatic and Social-Economic Factors influence on the Sucide Development in the Czech Republic. Paříž 22.08.2010 – 27.08.2010. In: COMPSTAT 2010. Paris: CNAM and INRIA, pp. 721–728. ISBN 978-3-7908-2603-6.

Bartošová, J. (2009). Analysis and Modelling of Financial Power of Czech Households. In: 8th International Conference APLIMAT 2009, Bratislava, February 3 – 6, 2009, Proceedings. Slovak University of Technology, Bratislava, pp. 717–722, ISBN 978-80-89313-31-0.

Bartošová, J., Bína, V. (2009). Modelling of income distribution of Czech households in years 1996 – 2005. Acta Oeconomica Pragensia, Vol. 17, No. 4/2009, pp. 3–18. ISSN 0572-3043.

Bartošová, J., 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, February 1–4, 2011, Proceedings. Slovak University of Technology, Bratislava, pp. 1451–1460, ISBN 978-80-89313-51-8.

Bílková, D. (2011). Use of the L-Moments Method in Modeling the Wage Distribution. Bratislava 01.02.2011 – 04.02.2011. In: APLIMAT 2011 [CD-ROM]. Bratislava: Slovak University of Technology, pp. 1471–1481. ISBN 978-80-89313-52-5.

Bílková, D. (2008). Modelování mzdových rozdělení v ČR v letech 2004 a 2005 s využitím logaritmicko-normálních křivek a křivek Pearsonova a Johnsonova systému. Statistika, Vol. 88, No. 2, pp. 149–166. ISSN 0322-788x.

Bílková, D., Budinský, P., Vohánka, V. (2009). Pravděpodobnost a statistika. 1. vyd. Plzeň: Vydavatelství Aleš Čeněk, 639 p. ISBN 978-80-7380-224-0.

Bílková, D., Malá, I. (2010). Development of income distributions of households in the Czech Republic since 1992. Praha 30.08.2010 – 03.09.2010. In: Prague Stochastics. Praha: KSTP VŠE.

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.

Miskolczi, M., Langhamrová, J. (2011). Labor market and simultaneous equations solved by TSLS. Bratislava 01.02.2011 – 04.02.2011. In: APLIMAT 2011 [CD-ROM]. Bratislava: Slovak University of Technology, pp. 1597–1608. ISBN 978-80-89313-51-8.

Pacáková, V., Sipková, L`., Sodomová, E. (2005). Štatistické modelovanie príjmov domácností v Slovenskej republike. Časopis pre ekonomickú teorii a hospodársku politiku, společensko-ekonomické prognózovanie (Journal of economics), Vol. 53, No. 4, pp. 427–439. ISSN 0013-3035.

Sipková, L`., Sodomová, E. (2007). Modelovanie kvantilovými funkciami. Bratislava: Vydavatel`stvo EKONÓM, 175 p. ISBN 978-80-225-2346-2.

Sipková, L`., Sodomová, E. (2009). Income distribution model in the Slovak Republic used the household SILC data. Studia i práce Uniwersytetu Ekonomicznego w Krakowie: Wspólczesne problemy modelowania i prognozowania zjawisk spoleczno-gospodarczynch, pp. 241–256. ISSN 1899-6205.

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

Ivana Malá

University of Economics in Prague

Faculty of Informatics and Statistics

Department of Statistics and Probability

nám. W. Churchilla 4, Praha, Czech Republic
malai@vse.cz