ESTIMATION OF CHANGES IN THE DISTRIBUTION OF INCOME IN THE CZECH REPUBLIC USING MIXTURE MODELS

Ivana Malá
University of Economics, Prague

ŚLĄSKI
PRZEGLĄD
STATYSTYCZNY
Nr 11 (17)

ISSN 1644-6739

Abstract: In the text the development of distributions (and their characteristics) of the equivalised net annual nominal income in Czech households in 2004-2009 is studied. Three-parametric lognormal and Dagum distributions are used as a model for income probability distribution. Moreover, finite mixtures of these distributions are estimated for the models with an observable component membership (given by the number of economically active members of the household and the number of unemployed members). Data from the European Union survey – Statistics on Income and Living Conditions 2005-2010 – are used for the analysis. All the estimates in the text are obtained using the maximum likelihood method.

Keywords: income distributions, maximum likelihood estimate, finite mixture, lognormal distribution, Dagum distribution.

1. Introduction

Studies and analyses in the field of incomes and wages are very important in the economy. Characteristics of their levels (as values of mean or median), characteristics of variability (standard deviation, coefficient of variation) and the Gini index of inequality are frequently published and discussed from various points of view. In this article, three-parametric lognormal and Dagum distributions [Dagum 1990; Kleiber 2008; Kleiber, Kotz 2003] and finite mixtures of these densities are used for the modelling of the probability distribution of the equivalised net annual income in the Czech Republic in the analysed period.

Three-parametric lognormal distribution is frequently used as a model of the probability distribution of incomes and wages as it is a positively skewed distribution [Cohen, Whitten 1980]. An exhaustive overview of the so called income distributions as generalized gamma, beta or lambda

distributions, Pareto or Weibull distributions can be found in [McDonald 1984] or [Kleiber, Kotz 2003]. The incomes (and wages) in the Czech Republic with the use of lognormal distribution are analysed in [Bartošová, Bína 2008; Bílková 2012; Bílková, Malá 2012] or [Pavelka 2009]. The last mentioned article − by Pavelka − shows the use of mixtures of lognormal distributions for wages in the Czech Republic.

In the article, the equivalised net annual income in the Czech Republic in the period 2005-2009 (in Czech koruna (CZK)) is analysed based on data from the European Union surveys – Statistics on Income and Living Conditions (EU-SILC). Dagum and lognormal distributions (and their mixtures) are used to model income distributions. The development of estimates of unknown parameters is of interest, as well as the estimated characteristics of location, variability or inequality. Moreover, the quality of fits is described and different fits are compared.

All unknown parameters are estimated with the use of the maximum likelihood method. If this method is used, it is easy to find the maximum likelihood estimates of parametric functions (as expected value, standard deviation or quantiles) substituting these estimates into parametric functions of interest.

2. Methods

In this text, two three-parametric income distributions are used. Detailed information about these distributions (and other income distributions) is given in [Kleiber, Kotz 2003]. For the lognormal distribution [Cohen, Whitten 1980; Kleiber, Kotz 2003], and for Dagum distribution results from [Dagum 1980; Kleiber 2007] were used.

Three-parametric lognormal distribution of a random variable X is described with the use of shift parameter θ, expected value μ of the logarithm of X and variance σ^2 of logarithm of X. This means that random variable $\ln(X-\theta)$ is distributed as $N(\mu,\sigma^2)$. The probability density function of the distribution is then of the form

$$f(x;\mu,\sigma^2,\theta)=\frac{1}{\sqrt{2\pi}\,\sigma(x-\theta)}\exp\left(-\frac{(\ln(x-\theta)-\mu)^2}{2\sigma^2}\right),\quad x>\theta. \quad (1)$$

The expected value $E(X)$ and percentiles x_P were computed with the use of formulas

$$E(X) = \theta + \exp(\mu + \sigma^2/2), \quad x_P = \theta + \exp\left(\mu + \sigma u_P\right) \quad 0 < P < 1, \tag{2}$$

where u_P is a $100P$ % quantile of standard normal distribution. The standard deviation of X does not depend on θ and it is given as

$$\sqrt{D(X)} = e^{\mu+\sigma^2/2}\sqrt{e^{\sigma^2} - 1}. \tag{3}$$

The Dagum distribution is a flexible distribution of the positive valued random variable that usually provides sufficient fit of income distribution. The density of the three-parametric Dagum distribution is given by the formula:

$$f(x;\alpha,\beta,p) = \frac{\alpha p x^{\alpha p-1}}{\beta^{\alpha p}\left[1+(x/\beta)^{\alpha}\right]^{p+1}}, \; x > 0, \tag{4}$$

where α, β and p are positive parameters. The distribution function corresponding to (4) can be written in the form

$$F(x) = \left(1+(x/\beta)^{-\alpha}\right)^{-p}. \tag{5}$$

Inverse function to F (quantile function) yields for $0 < P < 1$ to

$$x_P = F^{-1}(P) = \sqrt[\alpha]{\frac{\beta^{\alpha}}{\sqrt[p]{P^{-1}} - 1}}. \tag{6}$$

Expected value (for $\alpha > 2$) and variance were evaluated from the formulas

$$E(X) = \frac{\beta\,\Gamma\left(p + 1/\alpha\right)\Gamma\left(1 - 1/\alpha\right)}{\Gamma(p)},$$

$$D(X) = \frac{\beta^2\left(\Gamma(p)\Gamma\left(p + 2/\alpha\right)\Gamma\left(1 - 2/\alpha\right) - \Gamma^2\left(p + 1/\alpha\right)\Gamma^2\left(1 - 1/\alpha\right)\right)}{\Gamma^2(p)}, \tag{7}$$

where Γ is the gamma function. Formulas (7) are valid for $\alpha > 1$ for expected value and $\alpha > 2$ for variance.

It is known [Kleiber, Kotz 2003] that in the case of the two parametric lognormal distribution, the Gini coefficient depends only on the parameter σ. For the three parametric distribution this quantity depends on all parameters and can be evaluated as [Malá, Bílková 2012]:

$$G_{\text{lognormal}} = \frac{\exp\left(\mu+\sigma^2/2\right) erf\left(\sigma/2\right)}{\theta+\exp\left(\mu+\sigma^2/2\right)}, \tag{8}$$

where *erf* is the error function. The Gini coefficient for the Dagum distribution depends on shape parameters α and p and it is equal to [Kleiber 2008]:

$$G_{Dagum} = \frac{\Gamma(p)\Gamma(2p+1/\alpha)}{\Gamma(2p)\Gamma(p+1/\alpha)} - 1. \tag{9}$$

The finite mixtures (with K components), used in the text, have density function

$$f(x;\boldsymbol{\psi}) = \sum_{j=1}^{K} \pi_j f_j(x;\boldsymbol{\theta}_j), \tag{10}$$

where component densities f_j are three parametric lognormal or Dagum densities and weights π_j (mixing proportion) fulfil the condition $\sum_{j=1}^{K} \pi_j = 1,\ 0 \le \pi_j \le 1,\ j = 1, ..., K$. The vector of unknown parameters $\boldsymbol{\Psi}$ consists of (K-1) free parameters $\pi_j,\ j = 1, ..., K-1$ and K triplets $\boldsymbol{\theta}_j = \left(\mu_j, \sigma_j^2, \theta_j\right)$ for lognormal components, or $\boldsymbol{\theta}_j = \left(\alpha_j, \beta_j, p_j\right)$ for Dagum components. The expected value of a mixture is a mixture of component expected values, while the variances were evaluated from component variances and component expected values.

The estimation of unknown parameters is based on a sample $x_i, i = 1, ..., n$ with the size n. It is not possible to derive explicit formulas for the maximum likelihood estimates of unknown parameters of the distributions studied in this text, and numeric algorithms were used to maximize the logarithmic likelihood function.

In this text, we assume that for each observation group its component membership is observed. In this case, the problem of the estimation of unknown parameters can be split into components and parameters of component densities are estimated separately [Titterington, Smith, Makovet 1985]. For the estimation, the sample was divided into component subsamples with sample sizes n_j, $j = 1, ..., K$, $\sum_{j=1}^{K} n_j = n$. Maximum likelihood estimates $\hat{\pi}_j$ of mixing proportions, equal proportions of data from each component in the subsample $\hat{\pi}_j = \frac{n_j}{n}$, $j = 1, ..., K$. Maximum likelihood parameters for mixture components were found by numeric algorithms, as in the case of one fitted distribution mentioned above.

Both component distributions have three unknown parameters to be estimated. If only one distribution is used, the values of logarithmic likelihood function l in solutions can be compared. Mixture models with K components have $(K-1)+3K$ parameters. To compare all models with an unequal number of parameters (separately in years), Akaike's criterion was used.

All computations were made in the R [RPROGRAM...]. Parameters of lognormal distribution were evaluated according to [Cohen, Whitten 1980] and parameters of Dagum distribution were found using the package VGAM [RVGAM...].

3. Results

The survey EU-SILC (European Union – Statistics on Income and Living Conditions) has been performed by the Czech Statistical Office yearly since 2005 [CZSO 2012; EUROSTAT 2012]. The survey from 2005 deals with incomes in 2004, the last survey used in this text from 2010 covers incomes from 2009. The aim of the survey is to gather representative data on income distribution for the whole population and for various household types [CZSO 2012]. All incomes in the text are in CZK, the average yearly euro exchange rates and inflation rates in the Czech Republic are given in Figure 1. The equivalised net in-

come was evaluated for each household in the sample as a ratio of the total net income of a household, divided by the number of units that are constructed, to reflect the impact of sharing expenditure of members in the households. The weights were taken according to the EU scale as 1 for the first adult, 0.5 for other adults and 0.3 for each child in the household. It is clear that this income is greater than income per capita (and it is equal to it for single member households). Suppose that the equivalised income is the random variable X with the lognormal, Dagum (choice $K = 1$ for number of components) or mixture distribution with more components discussed in part 1. Among many EU-SILC variables, survey weights are also provided in order to eliminate the impact of the sampling procedure and to extend the sample to the set of households in the Czech Republic. This variable enables to recalculate data to the overall set of Czech households.

In the text both distributions were fitted into data. Then, mixture models with a known component membership were constructed for components defined by the number of economically active members (factor with 5 level – 0 to 3 and more than 3, $K = 5$) and number of unemployed members (factor with 3 levels – 0, 1, more than 1, $K = 3$). For 5 components there are 19 parameters ($4 + 5 \times 3$) and for 3 components there are 11 parameters ($2 + 3 \times 3$) to be estimated. The positive impact of the number of economically active members and the negative impact of number of unemployed members are intuitive and the aim of the analysis is to quantify this.

Sample sizes of EU-SILC surveys in the Czech Republic vary from 4,351 households in the first survey in 2005 to 11,294 households in 2008; all sample sizes are given in Table 1. The sample sizes in subpopulations differ from tens of households (for groups of households with more than three economically active members or more than one unemployed member) to thousands (households without unemployed members or households with 0-2 economically active members). Proportions of households for different components are given in Tables 4 and 5.

The study covers a six-year-long period. The development of the exchange rate of the Czech crown (koruna) and the inflation rate in the Czech Republic are shown in Figure 1. The inflation rate in the Czech Republic in the analysed period was 1.1413 [CZSO 2012].

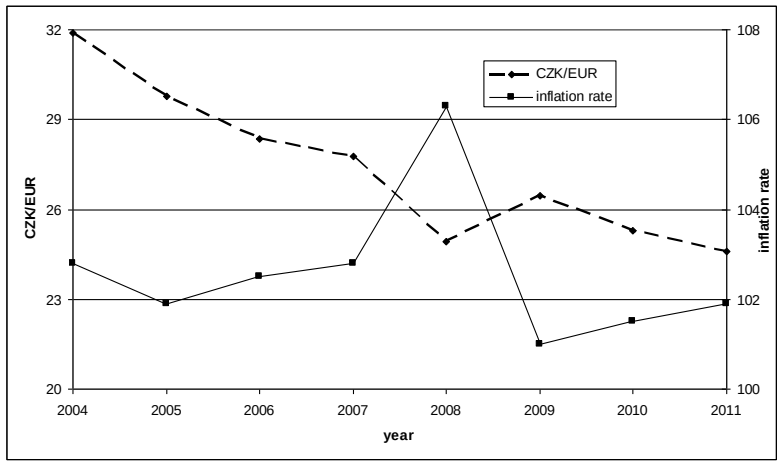

Figure 1. The development of the exchange rate of the Czech koruna and euro (left axis) and inflation rate in the Czech Republic (%, right axis)

Source: [CNB 2012; CZSO 2012].

In Table 1 the sample characteristics of location (average value, median) and absolute variability (standard deviation) are in CZK, the coefficient of variability and the Gini index are given as ratios. The growth in these characteristics from 2004 to 2009 was 35.88 % for the mean, 38.25% for the median and 24.37 % for the standard deviation. All these figures are greater than the inflation rate which was equal to 14.13 percent. The coefficient of variation decreased slightly by 5 percentage points as well as the sample Gini coefficient from 0.251 to 0.240.

Table 1. Sample characteristics of location and variability

Year	*n*	Average	Median	Standard error	Coeff. of variation	Gini coefficient
2004	4,351	148,261	127,500	94,052	0.634	0.251
2005	7,483	153,377	132,613	92,826	0.605	0.245
2006	9,675	165,468	143,548	93,689	0.566	0.240
2007	11,294	178,097	156,267	96,166	0.540	0.234
2008	9,911	193,878	169,120	119,103	0.614	0.239
2009	9,098	201,454	176,273	116,977	0.580	0.240

Source: own computations.

In Table 2 the estimated parameters for the lognormal and Dagum distributions fitted to data are shown. The rise of parameters μ is apparent, together with the stable value of estimates of σ. The small negative values for the shift parameter θ were found in 2007-2009. This fact frequently occurs when fitting lognormal distribution to income data. For the Dagum distribution, we can see high values and an increase in the scale parameter β, stable values of the shape parameter p and a slow increase in the shape parameter α.

Table 2. Estimated parameters for lognormal and Dagum distributions

Year	Lognormal distribution			Dagum distribution		
	μ	σ	θ	α	β	p
2004	11.800	0.436	66	3.778	113,069	1.499
2005	11.839	0.425	69	3.822	114,774	1.611
2006	11.919	0.419	41	3.897	126,230	1.560
2007	11.999	0.414	-180	4.026	139,875	1.475
2008	12.077	0.418	-22	3.977	149,714	1.514
2009	12.115	0.422	-92	4.013	160,259	1.395

Source: own computations.

From the estimates given in Table 2, the basic characteristics of the fitted distributions are not visible and for this reason maximum likelihood estimates, based on the estimated parameters from Table 2 of expected values, standard deviations and Gini coefficients are given for all the analysed years in Table 3. It can be seen that the maximum likelihood estimates of the characteristics are very similar from both fits and they correspond well with the observed values in Table 1, with the exception of the standard deviations. Estimates of standard deviations from both fits are lower than the observed values; however, the estimated standard deviation is greater for the Dagum model. As a measure of goodness-of-fit, the values of logarithmic likelihood function in the maximum likelihood solutions can be used. Lognormal distribution gives better fits for all the analysed years (values not given in the text), but no big difference was found. This is strange, as usually the Dagum distribution provides better fits to income data than the lognormal distribution.

Estimated expected values and standard deviations increased from 2004 to 2009 by 36% and 31 % for lognormal distribution, and by 35% and 26 % for the Dagum distribution. Estimates of the expected value are very similar for both fits; standard deviations are perceptibly greater for the Dagum distribution.

Table 3. Estimated expected value, standard deviation (CZK), the Gini coefficient for the fits from Table 1

	Lognormal distribution			Dagum distribution		
Year	expected value	standard deviation	Gini coefficient	expected value	standard deviation	Gini coefficient
2004	146,615	67,146	0.242	146,553	77,495	0.243
2005	151,728	64,441	0.236	151,676	78,157	0.237
2006	164,003	71,892	0.233	163,908	82,527	0.234
2007	176,893	76,477	0.230	176,437	85,592	0.228
2008	191,740	83,718	0.232	191,191	93,972	0.230
2009	199,630	88,373	0.235	198,798	97,653	0.231

Source: own computations.

These fits can be compared with the use of the mixture models mentioned above. If we use subgroups with a known subgroup membership, the fit usually does not improve much as in the case of artificial components constructed in order to improve the quality of the fit [Titterington, Smith, Makov 1985]. On the other hand, we obtained information about the distribution of incomes in components. If lognormal or Dagum distributions are mixed into a finite mixture, the result is not a lognormal or Dagum distribution. In this application of mixtures the components differ significantly in the expected values and the component variances are smaller than the overall variance; the components are more homogenous than the population. Estimated expected values of components for the mixture, according to the number of economically active members, are for both distributions and all the years shown in Figure 2. From this Figure the development of these characteristics is evident. The highest increase in the analysed income is in the last group with four or more economically active members (48% and 49 %). Other components have an increase of about 40 percent with the exception of the component with

three active members (only 27 % and 29 %). According to the AIC criterion, all these models are better than models with only one component, in the majority, lognormal fits are superior to Dagum. Remember, that estimates of mixing proportions are not dependent on component distributions and are the same for the lognormal and Dagum model.

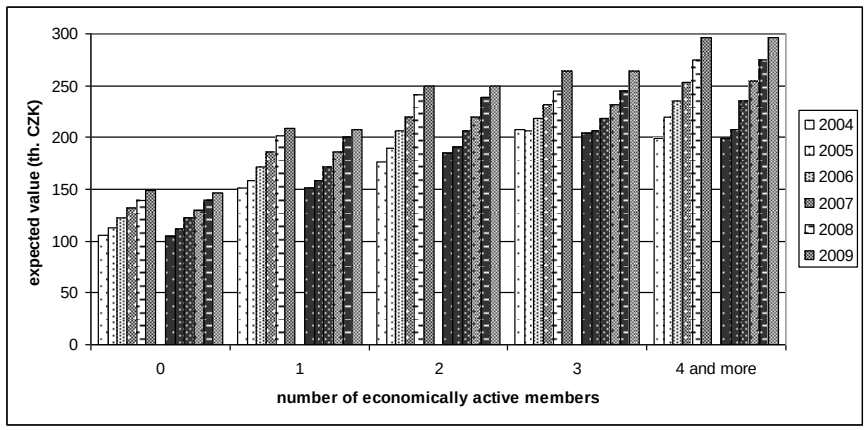

Figure 2. Estimated expected values for components according to number of economically active members, lognormal distribution (white, black), Dagum distribution (grey, white)

Source: own elaboration.

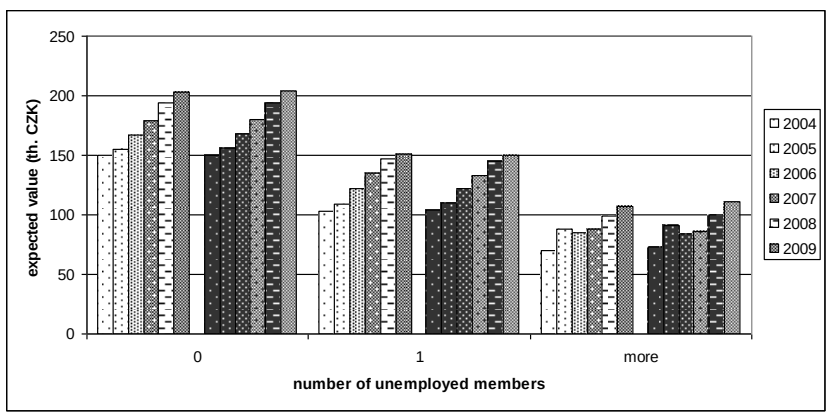

Figure 3. Estimated expected values for components according to number of unemployed members, lognormal distribution (white, black), Dagum distribution (grey, white)

Source: own elaboration.

In Figure 3, the estimated expected values are shown according to the number of unemployed members. The value for households with more than one unemployed member is less than half of the expected value for households without unemployed members. The increase in expected values from 2004 to 2009 for households without the unemployed, with one, and with more than one unemployed members is for lognormal distribution 35%, 47% and 52% (Dagum 35%, 41%, 52%). According to the AIC criterion, all these models are better than models with only one component and a majority of them is better than models with five components.

Table 4. Estimated mixing proportions and overall expected values and standard deviations

Year	Number of economically active members					Lognormal components		Dagum components	
	0	1	2	3	more	expected value	standard deviation	expected value	standard deviation
2004	0.320	0.303	0.306	0.054	0.017	148,509	62,063	150,751	79,526
2005	0.325	0.310	0.294	0.053	0.018	156,355	70,925	156,015	79,237
2006	0.325	0.310	0.294	0.053	0.018	169,074	75,688	168,731	83,088
2007	0.326	0.300	0.297	0.058	0.018	182,120	79,844	181,583	86,751
2008	0.318	0.305	0.298	0.061	0.017	197,190	87,359	196,550	92,832
2009	0.318	0.325	0.290	0.054	0.013	205,784	92,849	204,805	100,029

Source: own computations.

Table 5. Estimated mixing proportions and overall expected values and standard deviations

Year	Number of unemployed			Lognormal components		Dagum components	
	0	1	more	expected value	standard deviation	expected value	standard deviation
2004	0.870	0.110	0.020	143,330	65,992	143.908	79,225
2005	0.873	0.111	0.016	148,983	66,331	149,482	79,638
2006	0.884	0.096	0.020	161,086	70,731	161,576	84,525
2007	0.899	0.086	0.015	173,994	75,443	174,372	88,799
2008	0.915	0.074	0.011	189,081	82,994	189,376	97,318
2009	0.885	0.102	0.013	196,854	86,810	196,925	100,587

Source: own computations.

In Table 4 and Table 5, the estimated component proportions and overall expected values and standard deviations from the models with five components (Table 4) and three components (Table 5) are given. The characteristics of location and variability from these tables can be compared with the estimates in Table 3 and sample values in Table 1. The estimates of expected values in Table 5 (three components) are less than in Table 4 (five components).

4. Conclusions

In the text the net equivalised income in CZK in the period 2004-2009 in the Czech Republic is analysed. Two probability distributions frequently used with good results (lognormal, Dagum), and two mixture models of these distributions with known component membership are used. Both single distributions provide the acceptable model for the analysed incomes, however the goodness-of-fit tests are significant. Components defined by the number of economically active members and the number of unemployed members of the household define more homogenous subgroups of Czech households and provide better fits to income according to the AIC criterion.

In the analysis, information about components is obtained as a by-product. The development of components (component probabilities, expected values or other characteristics) can be analysed in a period of 6 years for the whole population.

The positive impact of the number of economically active members and the negative impact of the number of unemployed members are expected. In the text these relations were quantified with the use of the characteristics of the level and variability of equivalised income.

References

Bartošová J., Bína V., *Modelling of income distribution of czech households in years 1996-2005*, Acta Oeconomica Pragensia 2009, 17 (4), pp. 3-18.

Bílková D., Malá I., *Modelling the income distributions in the Czech Republic since 1992*, „Austrian Journal of Statistics" 2012, 41 (2), pp. 133-152.

Bílková D., *Recent development of the wage and income distribution in the Czech Republic*, "Prague Economic Papers" 2012, 21 (2), pp. 233-250.

CNB, Czech National Bank, www.cnb.cz. 10.10.2012.

Cohen A.C., Whitten J.B., *Estimation in the three-parameter lognormal distribution*, "Journal of American Statistical Association" 1980, 75, pp. 399-404.

CZSO, Czech Statistical Office, www.czso.cz, 10.10.2012.

Dagum C., *Generation and properties of income distribution functions*, [in:] *Income and Wealth Distribution, Inequality and Poverty: Proceedings of the Second International Conference on Income Distribution by Size: Generation, Distribution, Measurement and Applications*, 23-30 September 1989, University of Pavia, Italy, 1990, pp. 1-17.

EUROSTAT, http://epp.eurostat.ec.europa.eu/portal/page/portal/income_social_inclusion_living_conditions, 5.10.2012.

Kleiber C.A., *Guide to the Dagum Distributions*, working paper 23/07, Wirtschaftswissenechaftliches Zentrum (WWZ) der Univeritat Basel, URL, http://wwz.unibas.ch/uploads/tx_x4epublication/23_07.pdf. 2007.

Kleiber C.A., *Guide to the Dagum Distributions, Modeling Income Distributions and Lorenz Curves, Economic Studies in Equality, Social Exclusion and Well-Being*, Vol. 5, Springer New York 2008, pp. 97-117.

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

Mc Donald J.B., *Some generalized functions for the size distributions*, "Econometrica" 1984, 52 (3), pp. 647-663.

Pavelka R., *Application of density mixture in the probability model construction of wage distributions*, [in:] *Applications of Mathematics and Statistics in Economy: AMSE 2009*, Uherské Hradiště 2009, pp. 341-350.

RPROGRAM. R Core Team. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna 2012, http://www.R-project.org/.

RVGAM. Yee T. W. VGAM: Vector Generalized Linear and Additive Models, R package version 0.9-0. 2012, http://CRAN.R-project.org/package=VGAM.

Titterington D.M., Smith A.F.M., Makov U.E., *Statistical Analysis of Finite Mixture Distributions*, John Wiley, 1985.

SZACOWANIE ZMIAN W ROZKŁADZIE DOCHODÓW W CZECHACH Z WYKORZYSTANIEM MODELI MIESZANYCH

Streszczenie: W artykule przeprowadzono badanie rozkładu rocznych dochodów netto czeskich rodzin w latach 2004-2009. Jako model dochodów wykorzystano rozkład lognormalny Daguma. Ponadto uwzględniono skończone mieszanki rozkładów. Dane wykorzystane do analizy pochodzą z badania Unii Europejskiej: Statystyka dochodów i warunków życia 2005-2010. Wszystkie oszacowania parametrów otrzymano metodą największej wiarygodności.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY
Nr 11 (17)

Słowa kluczowe: rozkład dochodów, estymacja metodą największej wiarygodności, rozkład log-normlany, rozkład Daguma.