Proceedings of the 5th Symposium on Physics in Economics and Social Sciences, Warszawa, Poland, November 25–27, 2010

The Models of Personal Incomes in USA

P. Łukasiewicz[a], K. Karpio[a] and A. Orłowski[a,b]

[a]Katedra Informatyki SGGW, Nowoursynowska 159, 02-776 Warszawa, Poland

[b]Instytut Fizyki PAN, al. Lotników 32/46, 02-668 Warszawa, Poland

The shapes of distributions of personal incomes in USA have been investigated based on the data for 1993 to 2008. Comparisons between four models utilizing various number of parameters have been performed. The studies showed that the empirical data is described the best by the three-parameter Dagum model. Values of the models parameters indicate that the distribution of personal incomes can be regarded as zero-modal one. However, one-parameter exponential model shows a good agreement with data and can be treated as a good approximation of empirical distribution with the exception of the region with very high incomes. The high-income region is characterized by the relatively great number of events and is described much better by the Dagum distribution.

PACS: 89.65.-s, 89.65.Cd

1. Introduction

For a majority of countries we observe similar, characteristic shape of the income distribution. In the majority of societies we deal with the predominant number of entities (persons, families, households) of similar incomes as well as with a relatively small number of entities of high or very high incomes. That's why income distributions are single-modal with large right-sided asymmetry, additionally characterized by the "fat tail" in the range of very high incomes. Rarely we deal with zero-modal distributions. That is a case in poor countries, where a majority of entities gain incomes concentrated in a range of small incomes. However, in developed countries we may also obtain distributions of incomes similar to zero-modal ones. That takes place in the case of personal incomes in USA. In [1] authors approximated personal income distribution in USA with one-parametric exponential function given by the equation:

$$f_E(x) = \frac{1}{a}\exp(-\frac{x}{a}), \tag{1}$$

where x indicates individual income, whereas a parameter is equal to average income.

Model (1) is of course zero-modal. In this paper we investigate in more details a shape of the distribution of personal incomes in USA. We compare the results obtained in [1] with the results for three other models. The functions proposed have greater number of parameters, so they shall better compliance with empirical distributions. Moreover, depending on the shape of empirical distribution those models can became zero or one-modal. We are interested in answers to the following questions: (i) do the distribution of the personal incomes in USA can be regarded as zero-modal, or is it rather an one-modal distribution significantly "moved" towards small incomes (ii) do the models with greater number of parameters significantly improve agreement with empirical data (iii) can the exponential model be regarded as a good approximation of the income distribution in USA (iv) do the models proposed well describe distributions of very high incomes?

2. Models of incomes distributions

One of the directions of studies regarding incomes is search for mathematical functions which approximate empirical distributions. In literature there are proposals for various types of such functions. Some of them, as for example the Pareto model or log-normal and gamma distributions are currently rarely used (they have rather historical meaning). In some cases these functions can well approximate specific ranges of a distribution, e.g. incomes greater that a certain threshold. Very high accuracy with empirical distribution is characteristic for Dagum and Singh-Maddala models. These models are the three-parameter density functions with relatively simple analytical forms. There are also studies exploring the usefulness of some non-elementary functions, like beta distribution, generalized beta distribution, normal-Laplace distribution, generalized normal-Laplace distributions and others [2–5]. In this paper we use four models: exponential, Weibull, Dagum, and Singh-Maddala.

Density function of the two-parameter Weibull distribution has a form:

$$f_W(x) = \frac{b}{a}x^{b-1}\exp(-\frac{x^b}{a}), \tag{2}$$

where $x > 0$. Also a and b parameters are positive. For $b = 1$ this function reduces to (1). When $0 < b \leq 1$ the Weibull distribution is zero-modal. For $b > 1$ it is single-modal distribution. This model has been used, among others, during studies of incomes in [6]. Models (1) and (2) are characterized by the so-called "thin tails".

Density function of the Dagum distribution [7] is described by the equation

$$f_D(x) = \frac{abc}{x^{b+1}(1+ax^{-b})^{c+1}}, \tag{3}$$

where $x > 0$, and the parameters fulfill the following conditions: $a > 1, b > 0$, and $c > 0$. This distribution is zero-modal when $0 < bc < 1$ and single-modal for $bc > 1$.

The Singh-Maddala distribution [8] can be expressed in the form of

$$f_{SM}(x) = \frac{abcx^{b-1}}{(1+ax^b)^{c+1}}, \tag{4}$$

where $x > 0$, and $a > 0, b > 0, c > 0, bc > 1$. This

distribution is zero-modal when $0 < b < 1$ and single-modal for $b > 1$.

Studies performed in various countries show that models (3) and (4) exhibit high conformance to empirical distributions of incomes. Very often the Dagum model is utilized [9–12]. The other advantage of these functions is a small number of finite moments. Curves (3) and (4) have "fat tails" what is their advantage because empirical distributions are usually extended in the range of incomes exceeding average. Models (2), (3), and (4) are universal, they may describe zero-modal distributions as well as single-modal ones.

3. Income data and models estimation

Data analyzed in this paper contain information, among others, about personal incomes in USA in 1993 to 2008. Files with data have been downloaded from http://www.census.gov [14]. Data for 1993, 1996, 2004, and 2008 have been collected within the project "Survey of Income and Program Participation" (SIPP), whereas data for 1998, 2000 and 2002 within the "Annual Demographic Survey" (March CPS Supplement). The variable "total person's income" has been studied. Data undergone preliminary selection: zero values (lack of data) have been eliminated, monthly income has been recalculated into annual income. Incomes are expressed in k$ (thousand of dollars).

The a parameter's estimator of the model (1) is, of course, equal to arithmetic average from the sample. The parameters of the models (2), (3), and (4) have been evaluated with the means of the maximum likelihood method. Data have been grouped in 1800 bins with the width 0.6 k$ each. Vector $\hat{\theta}$ of the values of a model's parameters have been calculated by maximizing the function:

$$\ln(L(\theta)) = \ln\left(\frac{n!}{n_1!n_2!\ldots n_{1800}!}\prod_{i=1}^{1800}(p_i(\theta))^{n_i}\right) = \\ = \text{const} + \sum_{i=1}^{1800} n_i \ln(F(c_i;\theta) - F(c_{i-1};\theta)), \quad (5)$$

where F is a cumulative distribution function (CDF) of the theoretical distribution, θ is the vector of parameters being evaluated, c_{i-1}, c_i are the lower and upper edges of the i-th bin, n_i is a number of incomes in an i-th bin, and $n = n_1 + n_2 + \ldots + n_{1800}$.

For graphical presentation as well as to determine measures of compliance between theoretical and empirical distributions data have been grouped in 300 bins with the width 10/3 k$ each (the same way as in [1]). For each evaluated model the sum of squared residuals (the sum of squared errors or SSE) and the sum of absolute values of the residuals (the sum of absolute errors or SAE) are calculated, according to the equations

$$SSE = \sum_{i=1}^{300}[\frac{n_i}{n} - p_i(\hat{\theta})]^2,$$

$$SAE = \sum_{i=1}^{300}|\frac{n_i}{n} - p_i(\hat{\theta})|. \quad (6)$$

In addition, two other (but related) measures are computed:

$$W_r = (1 - \frac{1}{2}SAE) \times 100\%,$$

$$W_p = \sum_{i=1}^{300}\min(\frac{n_i}{n}, p_i(\hat{\theta})) \times 100\%. \quad (7)$$

In the case of correct choice of a theoretical distribution we shall observe good compatibility between values of empirical c_i and theoretical q_i quantiles. As the measure of such a compatibility a correlation coefficient squared ρ^2 between quantiles c_i and q_i has been used [13]. Correlation coefficients ρ have been calculated based on the 199 quantiles of $i/200$ rank (i = 1, 2, ..., 199), evaluated on the basis of individual data.

4. Shapes of income distributions

The results of the models estimation are presented in the Table. For each year the best compatibility is obtained for Dagum model. It yields the lowest values of SSE and SAE as well as the highest values of W_r, W_p, and ρ^2, thus being the best choice with respect to all quality measures. For years from 1996 to 2000 the second model, judged by the level of compatibility, is the Singh-Maddala model, while the third one is the exponential. Starting from 2002 we observe the inverse order between the two latter models but the differences are small. For all years except 1993, the worst agreement between the model and the empirical distributions is observed for the Weibull distribution. For each investigated year the estimated Dagum models are always the zero-modal distributions ($\hat{b} \cdot \hat{c} < 1$), the estimated Singh-Maddala models are always the single-modal distributions ($\hat{b} > 1$), and the estimated Weibull models are always zero-modal distributions ($\hat{b} \leq 1$). Plots of the density functions for 1996 and 2008 are presented in Fig. 1.

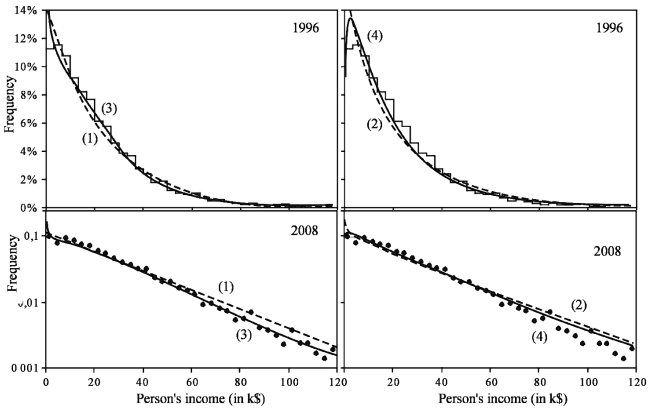

Fig. 1. Models of income distributions for 1996 and 2008. There are logarithmic scales on the bottom plots. Models: (1) Exponential (2) Weibull (3) Dagum (4) Singh-Maddala.

TABLE

Results of the fits of income models.

Year	Distribution	Parameters							
Count		$\hat{a}$	$\hat{b}$	$\hat{c}$	SSE	SAE	W_r	W_p	ρ^2
2008	Exponential	30.163	–	–	0.00079	0.12463	93.77	93.74	0.9075
	Weibull	22.896	0.9284	–	0.00176	0.17019	91.49	91.27	0.9159
288,454	Dagum	32,897.3	2.7086	0.2964	0.00071	0.09969	95.02	94.50	0.9479
	Singh-Maddala	0.0048	1.0378	6.9577	0.00108	0.12613	93.69	93.74	0.9323
2004	Exponential	26.440	–	–	0.00048	0.11134	94.43	94.40	0.9729
	Weibull	19.056	0.9123	–	0.00185	0.17334	91.33	91.01	0.9785
203,231	Dagum	33,598.4	2.7982	0.2819	0.00040	0.06787	96.61	95.97	0.9924
	Singh-Maddala	0.0062	1.0453	6.1825	0.00091	0.12492	93.75	93.81	0.9871
2002	Exponential	31.879	–	–	0.00061	0.12733	93.63	93.61	0.9075
	Weibull	24.178	0.9299	–	0.00133	0.17110	91.45	91.24	0.9157
145,209	Dagum	29,473.9	2.6752	0.3148	0.00024	0.08369	95.82	95.46	0.9495
	Singh-Maddala	0.0055	1.0847	5.1106	0.00084	0.13071	93.46	93.58	0.9371
2000	Exponential	28.591	–	–	0.00075	0.12677	93.66	93.63	0.9843
	Weibull	26.535	0.9806	–	0.00101	0.13917	93.04	92.97	0.9856
93,970	Dagum	29,473.9	2.7434	0.3183	0.00027	0.07094	96.45	96.17	0.9987
	Singh-Maddala	0.0053	1.1228	5.1245	0.00046	0.08999	95.50	95.67	0.9956
1998	Exponential	26.712	–	–	0.00076	0.12799	93.60	93.57	0.9827
	Weibull	21.953	0.9480	–	0.00159	0.16335	91.83	91.65	0.9859
92,080	Dagum	29,384.0	2.7973	0.3089	0.00036	0.07029	96.49	96.16	0.9976
	Singh-Maddala	0.0068	1.1333	4.3675	0.00049	0.09885	95.06	95.26	0.9957
1996	Exponential	23.611	–	–	0.00087	0.11869	94.07	94.02	0.9901
	Weibull	18.726	0.9363	–	0.00214	0.16545	91.73	91.47	0.9919
256,408	Dagum	22,020.6	2.7926	0.2977	0.00055	0.08041	95.98	95.49	0.9995
	Singh-Maddala	0.0070	1.0971	5.2941	0.00083	0.11057	94.47	94.64	0.9964
1993	Exponential	17.330	–	–	0.0021	0.1285	93.57	93.50	0.9885
	Weibull	18.624	1.0215	–	0.0016	0.1148	94.26	94.26	0.9874
142,162	Dagum	26,359.9	3.0252	0.2719	0.0011	0.1047	94.76	94.13	0.9966
	Singh-Maddala	0.0022	1.0499	23.4414	0.0014	0.1091	94.54	94.62	0.9902

For all empirical distributions the existence of relatively large number of very high incomes is observed. In Fig. 2 we present the best fits of all four models (exponential, Weibull, Dagum, and Singh-Maddala) to the empirical distribution of incomes for 2004. Part b) of Fig. 2 is an amplification of the high incomes region. In Fig. 3 the same models are fitted to the empirical distributions of incomes for 1998 and 2004. This time the logarithmic scale is used to highlight differences in the models quality as applied to the high incomes regions. It is clear that the Dagum model is the best approximation to the empirical distributions. The exponential model does not explain large number of events observed in the high incomes region.

It is intersting to note that the shape of the investigated empirical distributions of incomes exhibit a rather peculiar and irregular behavior in the region of high incomes. For each year there is a range of (high) incomes that is more populated than its neighborhood from both sides. For example, in 1998 there were about 440 observations within range of 330–480 k$, in 2004 about 780 observations were registered within range of 380–500 k$), while in 2008 about 1800 observations fell into 300–420 k$ interval. This effect seems not to be properly understood and suggests that more detailed analysis of the high incomes region is desirable. We plan to investigated this and related issues in a forthcoming paper.

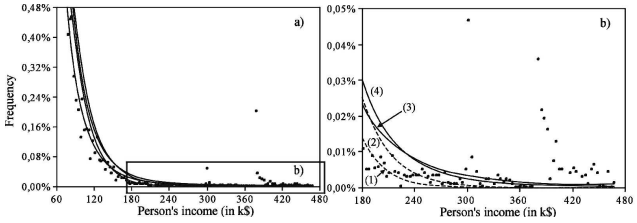

Fig. 2. Fits to incomes distribution for 2004. High incomes region is magnified in part b). Models: (1) Exponential (2) Weibull (3) Dagum (4) Singh-Maddala.

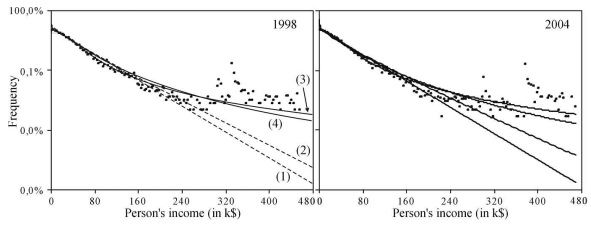

Fig. 3. Fits to incomes distributions for 1998 and 2004. Logarithmic scale is used to better resolve the fit quality for the high incomes regions. Models: (1) Exponential (2) Weibull (3) Dagum (4) Singh-Maddala.

5. Final conclusions

1. The best agreement with empirical distributions is observed for the Dagum model. For all the investigated years the zero-modal distributions are obtained.
2. The Singh-Maddala model is characterized by the lower level of compatibility with empirical data than the Dagum model. For all the investigated years the single-modal distributions are obtained.
3. The exponential model can be considered as a good approximation of personal income distributions but only up to a certain threshold value of income.
4. The exponential model does not explain the incomes behavior in the high incomes range. In that region the best approximation is provided by the Dagum model.
5. The quality of the exponential model (being the special case of the two-parametric Weibull model) cannot be improved by incorporating a second parameter.

References

[1] A. Drăgulescu, V. M. Yakovenko, Evidence for the exponential distribution of income in the USA, *The European Physical Journal B* **20**, 585-589 (2000).

[2] J.B. McDonald, Y.J. Xu, A generalization of the beta distribution with applications, *Journal of Econometrics* **66**, 133-152 (1995).

[3] J.B. McDonald, Some Generalized Functions for the Size Distribution of Income, *Econometrica* **52** (3), 647-663 (1984).

[4] Fan Wu, *Applications of The Normal Laplace and Generalized Normal Laplace Distributions*, A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in the Department of Mathematics and Statistics, University of Victoria, 2008.

[5] M. Jagielski, R. Kutner, Study of Households' Income in Poland by Using the Statistical Physics Approach, *Acta Physica Polonica A* **117** (4), 615-618 (2010).

[6] C.P.A. Bartels, H. van Metelen, *Alternative Probability Density Functions of Income*, Vrije University Amerstadam: Research memorandum 29, 1975.

[7] C. Dagum, A New Model of Personal Income Distribution: Specification and Estimation, *Economic Appliquee*, **XXX** (3), 413-437 (1977).

[8] S.K. Singh, G.S. Maddala, A Function for Size Distribution of Incomes, *Econometrica* **44** (5), 963-970 (1976).

[9] R. Bandourian, J.B. McDonald, R.S. Turley, *A Comparison of Parametric Models of Income Distribution Across Countries and Over Time*, Luxembourg Income Study Working Paper **305**, 2002.

[10] C. Dagum, A. Lemmi, *A Contribution to the Analysis of Income Distribution and Income Inequality and a Case Study: Italy*, Econometric Society Meeting, Chicago, 1987.

[11] P. Łukasiewicz, A.J. Orłowski, Probabilistic Models of Income Distributions of Polish Households, *Quantitative Methods in Economic Research III*, 122-130 (2003) [in Polish].

[12] P. Łukasiewicz, A.J. Orłowski, Probabilistic Models of Income Distributions, *Physica A* **344** (1-2), 146-151 (2004).

[13] R.B. D'Agostino, M.A. Stephens, Goodness-of-fit Techniques, Marcel Dekker, Inc., New York, 1986.

[14] U.S. Census Bureau, http://www.census.gov .