Proceedings of the 10th Polish Symposium on Physics in Economy and Social Sciences (FENS 2019)

Three-Part Model of Distribution of Household Incomes

P. Łukasiewicz*, K. Karpio and A.J. Orłowski

Institute of Information Technology, WULS-SGGW, Nowoursynowska 159, 02-776 Warsaw, Poland

Doi: 10.12693/APhysPolA.138.96 *e-mail: piotr_lukasiewicz@sggw.edu.pl

Previous studies of Polish household budgets usually assumed two-part distributions of the incomes. Low and medium incomes have been approximated by a log-normal distribution. On the other hand high incomes have been described by the Pareto Type I model. In this paper we analyze cumulative distribution of Polish household incomes. We observe significant deviations from the log-normal model describing the lowest incomes. We propose a model which better describes the left end of the income distribution. Eventually we construct a distribution consisting of the three parts, which seems to be well suited to correctly describe the whole range of the incomes. We also investigate non-positive incomes and the relation between them and expenditures.

topics: tails of income distribution, power law, log-normal model

1. Introduction

Studies of incomes conducted within two last decades show characteristic and universal for a majority of countries shape of income distribution. The distribution is approximately log-normal with deviation from it in the upper tail. Above the certain income the distribution is dominated by a power-law. The upper tail is occupied by a small number of objects (about 1%–2%), but they own a large fraction of total sum of incomes. This result has been known since the studies of V. Pareto [1] and dynamics of formation of the right tail of the distribution is described in literature [2–4]. In turn, a log-normal shape for low and medium incomes is related historically with the Gibrat's law of proportionate effect [5]. However, the assumptions in Gibrat's model are in disagreement with empirical evidence [6]. Additionally, Kalecki [7] noted that Gibrat's model leads to the variance of the distribution which increases indefinitely with time. Despite of many modifications of the Gibrat's model a dynamics of the log-normal shape for low and medium incomes has not been sufficiently explained. It is worth noting that this issue was explained for personal incomes in USA and Japan, where the income distribution up to the about 90th percentile has an exponential shape [8].

The joint log-normal & power law model was empirically confirmed for personal as well as household incomes in many countries [9–11]. The same results were also obtained for Polish households [12] for years 2003 and 2006. In [13] we compared the log-normal & power law model with other models of incomes for years 2004–2012 and we confirmed earlier results.

Analyzing cumulative distribution of Polish household incomes, we observe significant deviations from the log-normal model for the lowest incomes. The range of the lowest incomes is narrow (about 1%) but it is very important from the socio-economic point of view. This part of the income distribution is the subject of the studies of poverty. In this paper we propose a model which well describes the left end of the income distribution. We also take into account non-positive incomes which allow us to better understand a shape of the left tail.

2. Data and models evaluation

Data from the Household Budget Survey (HBS) project for years 2000–2015 was used in this work. We studied total available income of households based on the data contained about 32,000–37,000 of households depending on year. Household's available income is a sum of household's gross incomes from various sources reduced by all income taxes as well as by the social security and health insurance taxes. The available income comprises: wages and salaries, incomes from farms, self-employment, properties, rents, various social benefits (including retirement pensions and pensions), and other incomes (e.g. alimonies). Available income is allocated to expenditures and savings increase.

As in previous studies [12], in order to compare our results, we studied total incomes of the household instead of dividing it per number of persons. We recalculated registered monthly incomes into the annual ones.

The models evaluation was performed based on the empirical cumulative distribution $F_{\text{emp}}(x_j) = k_j/N$, where data x_j for $j = 1, \ldots, N$ are sorted ascending and k_j is rank of income x_j.

The theoretical cumulative distributions fitted to data have the following forms: (i) Pareto cdf: $F(x;\alpha) = 1 - (x_M/x)^{\alpha}$ for $x \geq x_M$, where $\alpha > 0$ is Pareto exponent; (ii) log-normal cdf: $F(x;\mu,\sigma) = \Phi((\ln(x) - \mu)/\sigma)$ for $x_m < x < x_M$, where Φ is a cdf of standard normal distribution, and μ, $\sigma > 0$ are parameters, interpreted as the mean and standard deviation of the logarithm of income. In order to describe the left tail of the empirical distribution we propose the power model:

$$F(x;\beta) = \left(\frac{x}{x_m}\right)^{\beta}, \tag{1}$$

for $0 < x \leq x_m$, where $\beta > 0$ is a parameter. The density function of this distribution has a form:

$$f(x;\beta) = \left(\frac{\beta}{x_m}\right)\left(\frac{x}{x_m}\right)^{\beta-1}. \tag{2}$$

We fit the power functions to data using the standard method, transforming them to the linear form. The Pareto model we fit to the complementary cumulative distribution $1 - F_{\text{emp}}$. Starting from the preliminary range of data with each iteration we extend this range and measure the root mean squared error,

$$\text{RMSE} = \frac{1}{k}\sqrt{\sum_{j=1}^{k}\left[F(x_j) - F_{\text{emp}}(x_j)\right]^2}.$$

The final limit values x_m and x_M were set when the RMSE started to systematically grow up. The log-normal function was fitted to data in the range $x_m < x < x_M$ using the nonlinear last squared method.

3. Three parts of the income distribution

We evaluated models for each of the three parts of the income distribution using the described above method. The analysis covered all years within the range from 2000 to 2015. We observe the common rule for all the analyzed years. The highest incomes are distributed according to the power law, middle and low incomes are well described by the log-normal distribution, while the lowest incomes follow the power model. The tails of the distribution of Polish household incomes are both described by the power functions. The domain of each model varies with years and do not exhibit any visible regularities. Each part of the income distribution, described by the corresponding model, covers the following part of the population: (1) the right tail 14% ÷ 26%, (2) the main part of the distribution 73% ÷ 85%, (3) the left tail 0.6% ÷ 1.2%.

We would like to pay attention, that a tail of the distribution is a term which is not precisely defined in statistics. Usually a tail refers to the part of the distribution which is really far away from the mean value. On the other hand a beginning of a tail can be referred as the place where the distribution changes its shape. In this work we estimated the limit values x_m and x_M (for left and right tail) by fitting appropriate models to data.

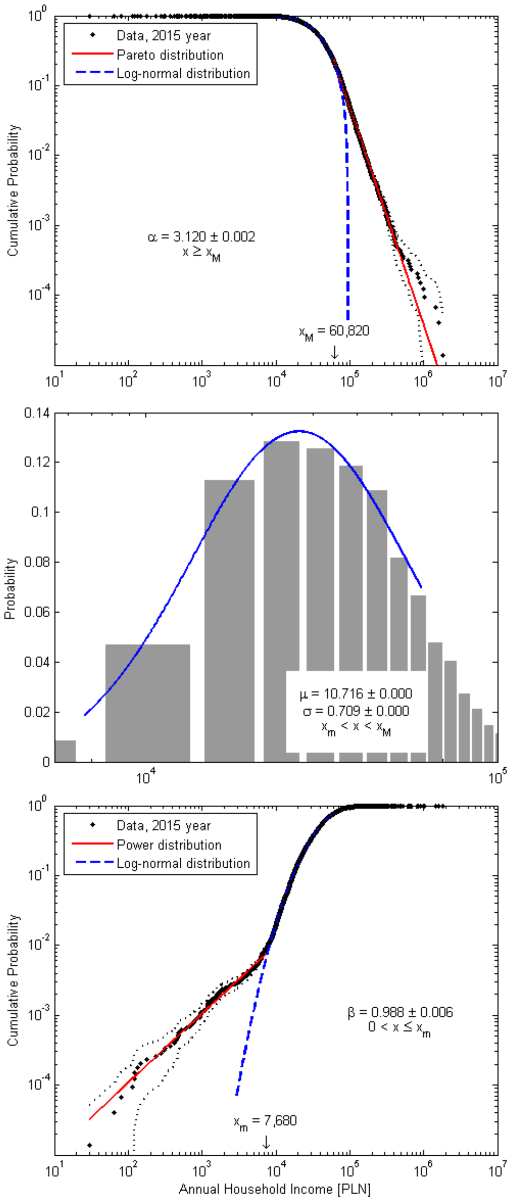

Fig. 1. Three parts of the household income distribution in 2015. Top plot: right tail of the empirical complementary cumulative distribution and Pareto model (solid line) in log-log scale. Middle plot: histogram for the main part of the income distribution and log-normal model in semi-log scale. Bottom plot: left tail of the empirical cumulative distribution and power model (solid line) in log-log scale. Dotted line: 95% confidence interval of the empirical cdf, dashed line: log-normal model outside the range of its estimation.

The results of the models' estimations for year 2015 are summarized in Fig. 1. The complementary cumulative distribution is shown in the log-log

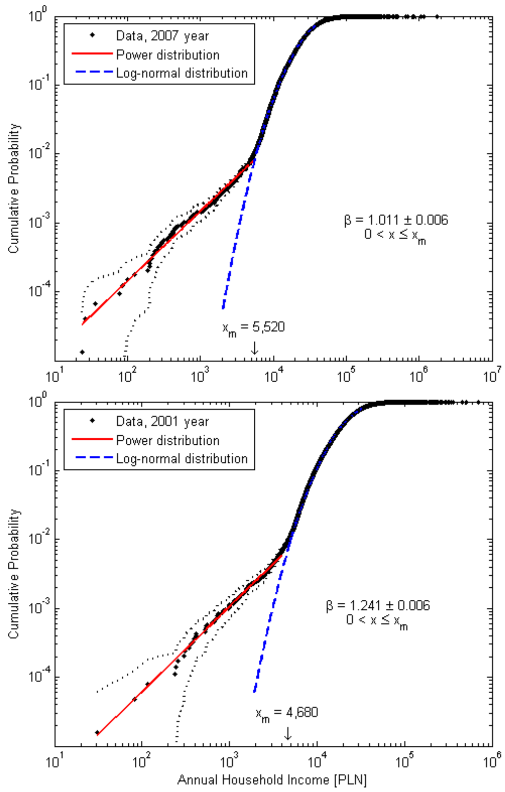

Fig. 2. Left tail of the empirical cumulative distribution in 2001 and 2007 in log-log scale. Solid line: power model, dotted line: 95% confidence interval of the empirical cdf, dashed lines: log-normal model outside the range of its estimation.

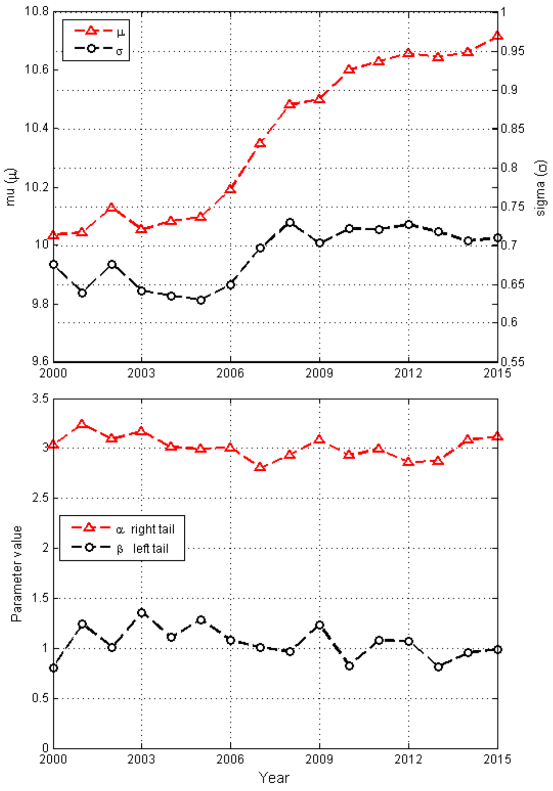

Fig. 3. Estimators of the models parameters for 2000–2015: $\hat{\mu}$, $\hat{\sigma}$ – log-normal model, $\hat{\alpha}$ — Pareto exponent, and $\hat{\beta}$ — power model. Error bars are the size of the data points.

scale in Fig. 1a. Dotted lines are 95% confidence interval limits of the empirical cdf estimated based on Wald method [14]

$$F_{\text{emp}}(x_j) \pm 2\sqrt{\frac{F_{\text{emp}}(x_j)}{N}\Big(1 - F_{\text{emp}}(x_j)\Big)}.$$

The Pareto distribution is represented by a straight line with $-\alpha$ slope. The power law model describes the income distribution above 60,820 PLN (about 22% of the investigated population). In Fig. 1b, the main part of the income distribution is presented as the histogram with the density function of the log-normal model. The range of incomes described by the log-normal model is from 7,680 PLN to 60,820 PLN (about 77% of the population). Figure 1c contains the low part of the income distribution below 7,680 PLN (about 1% of the population). The power function is represented, in the log-log scale, by the strength line with the slope β. The model is highly compliant with the data, the observed deviations do not reach one third of the confidence interval. A similar agreement with data is observed for all the years. The sample of the results is presented in Fig. 2 for years 2001 and 2007.

In the next step of the analysis we investigated values of the models parameters estimators in time. They are presented as $\hat{\mu}$, $\hat{\sigma}$, $\hat{\alpha}$, and $\hat{\beta}$ in Fig. 3. Their errors are of the size of data points thus they were not being indicated. Increasing values of $\hat{\mu}$ reflect the movement of the main part of the income distribution towards the higher incomes. They show the increase of the mean logarithm of incomes in the group of households with low and middle incomes. The clearly visible stability of the remaining parameters is very interesting. The $\hat{\sigma}$ parameter reflects the Gaussian width of the main part of the income distribution in a semi-log scale. The increase of the $\hat{\sigma}$ after 2005 is followed by its stability at about 0.72 in years 2007–2015. The Pareto exponent $\hat{\alpha}$ characterizing the right tail of the distribution changes between the limits $2.81 \div 3.24$ and fluctuates around 3.00. Similarly, an exponent $\hat{\beta}$ describing the left tail fluctuates within $0.81 \div 1.36$ and after 2006 it stabilizes at about 1.00. The unit value of the parameter corresponds to an uniform distribution of incomes.

4. Structure of the left tail of the income distribution

The results obtained for the left tail indicate flat distribution: a number of households with certain income does not depends on the income value. Incomes of the households in the left tail are random in the range of $0 \div x_m$. This phenomenon

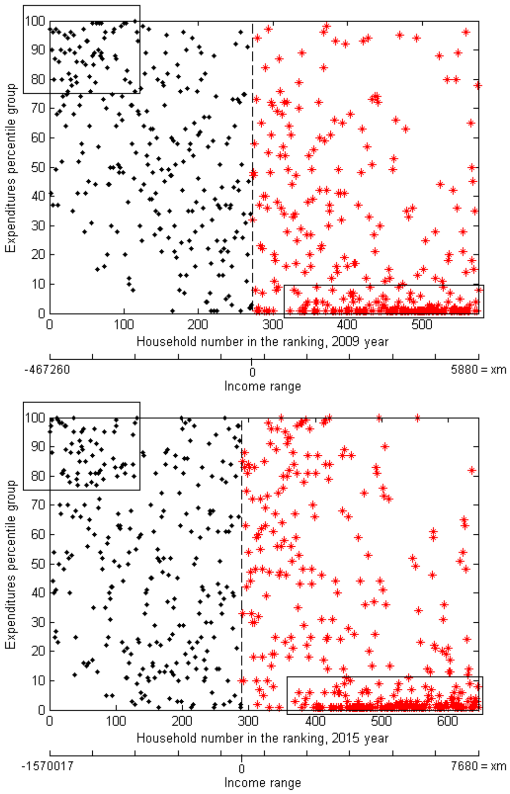

Fig. 4. Expenditures vs. incomes for $x \leq x_m$. Expenditures are showed as the percentile groups. Incomes are expressed as the income rank starting from 1 for the lowest income. The range of income values is indicated on the additional horizontal axis. The plots contain data for 2009 (top) and 2015 (bottom). The vertical dashed line separates non-positive incomes (dots) and left tail incomes (stars). See text for details.

is not usual in the income distributions and it may indicate we deal with the households which are not typical. In order to further investigate these households we compare incomes with expenditures. For completeness, we also take into account households with non-positive incomes. These households gain at least part of their total incomes from business activities. They report a lack of available incomes (incomes equal to zero) or a loss (negative incomes) in the survey studies. The number of such a households is approximately equal to the number of households in the left tail. Dispersion of expenditures vs. income rank is presented in Fig. 4 for households with incomes $x \leq x_m$. The plots contain data for 2009 (top) and 2015 (bottom). Because of the very wide range of income values, we presented the rank of incomes on the horizontal axis instead of income values. The rank starts from 1 for the minimal income. The range of income values is indicated on the additional horizontal axis. Similarly, the expenditure percentile group number instead of expenditures values is given on the vertical axis. Thus, each point in Fig. 4 has two integer coordinates: position in the income ranking and expenditure percentile group number. The vertical dashed line splits data into two income groups: non-positive and left tail incomes. Additionally, data points in both groups are indicated in different way.

One shall expect an increase of expenditures with income rank. However, this behavior is observed only for some of households. Those households are clearly visible in the right — low part of Fig. 4 for 2015, as the concentration of data points (indicated by rectangle). The expenditures for those households are in the lowest percentile groups. This is the first group of the households with low incomes, as well as expenditures.

Let note the remaining data points in parts of the plots for positive incomes are predominantly distributed uniformly. The similar distribution of data points is observed for the majority of the non-positive incomes (except some excess of the high percentile groups for the lowest incomes, indicated by rectangle). The observed similarity indicates the both groups could be considered together. This is the observed second group of households. For these households expenditures and incomes are not related to each other. These households declare low positive or non-positive incomes, while their expenditures are observed in all percentile groups. The possible explanation of the observed behavior is that these households declare high costs of running business while obtaining high revenues. The amount of their expenditures indicates on their real incomes hidden in the business expenses.

The observed low incomes in the left tail of distribution are the results of the balances between the revenues and the costs. The households can declare low incomes because of the true low revenues or the high revenues and the high costs. For the former ones the expenditures are also small, while for the latter ones observed low incomes are random because they are much smaller that the revenues and costs. In other words, reporting low incomes having high revenues requires a declaration of high costs, e.g. in order to avoid taxes. Expected difference revenues-costs is not necessary equal to zero, it can be negative, zero or small positive. This explains flat distribution of the lowest incomes.

5. Summary

In this paper we analyze cumulative distribution of Polish household incomes in years 2000–2015. We distinguish three parts of the income distribution which we parametrize using dedicated models. Based on the previous findings we approximate high incomes by the Pareto Type I model, low and medium incomes by log-normal model. For the lowest incomes we propose a power model. The values of the models parameters are evaluated for each year and their changes discussed. A huge range (covering on average 20% of the population) of the right tail and its stable slope (about 3.0) are characteristic for Polish household incomes. The width of the main

part of the distribution is also stable (about 0.72) after 2006. The slope of the left tail is stable (about 1.0, since 2007) too. The observed slope is compatible with the uniform flat distribution of incomes. Taking into account non-positive incomes, as well as household expenditures we explain the shape of the left tail as the result of the balance between the revenues and the costs.

References

[1] V. Pareto, *Cours d'Economie Politique*, F. Rouge, Lausanne 1896–97.

[2] D.C. Manrubia, D.H. Zanette, *Phys. Rev. E* **59(5)**, 4945 (1999).

[3] X. Gabaix, *J. Econ.* **114**, 739 (1999).

[4] M. Nirei, W. Souma, in: *The Complex Dynamics of Economic Interaction*, Eds. M. Gallegati, A.P. Kirman, M. Marsili, Springer, Berlin Heidelberg 2004, p. 161.

[5] R. Gibrat, *Les Inégalités économiques*, Paris 1931.

[6] Y. Fujiwara, C. Di Guilmi, H. Aoyama, M. Gallegati, W. Souma, *Physica A* **335**, 197 (2004).

[7] M. Kalecki, *Econometrica* **13**, 161 (1945).

[8] M. Nirei, W. Souma, *Rev. Income Wealth* **53(3)**, 440 (2007).

[9] W. Suoma, *Fractals* **09**, 463 (2001).

[10] F. Clementi, M. Gallegati, in: *Econophysics of Wealth Distributions*, Eds. A. Chatterjee, S. Yarlagadda, B.K. Chakrabarti, Springer, Milano 2005.

[11] F. Clementi, M. Gallegati, *Physica A* **350**, 427 (2005).

[12] M. Jagielski, R. Kutner, *Acta Phys. Pol. A* **117**, 615 (2010).

[13] P. Łukasiewicz, K. Karpio, A. Orłowski, *Equilibrium* **13**, 603 (2018).

[14] A. Agresti, B.A. Coull, *Am. Stat.* **52(2)**, 119 (1988).