

Alina Jędrzejczak

Uniwersytet Łódzki

UWAGI O SZACOWANIU PARAMETRÓW ROZKŁADU DOCHODÓW

1. Wstęp

Dane na temat rozkładów płac i dochodów w Polsce pochodzą głównie z badań częściowych (budżetów gospodarstw domowych lub badania aktywności ekonomicznej ludności). Badania te są często krytykowane ze względu na różnego rodzaju błędy o charakterze statystycznym i pozastatystycznym. Poza błędami problemem dla badacza może być także forma, w jakiej dane na temat płac i dochodów są udostępniane – najczęściej dysponujemy danymi pogrupowanymi i jednocześnie uciętymi ze względu na brak odpowiednich informacji w próbie lub jej pominięcie. Innym problemem, z którym szczególnie często spotykamy się w przypadku badania płac i dochodów, są informacje nietypowe związane z występowaniem dochodów bardzo niskich lub bardzo wysokich. Wszystko to skłania do poszukiwania odpornych metod wnioskowania statystycznego, które mogą być wykorzystane do uzupełnienia metod klasycznych. W pracy przedstawiono odporne metody estymacji, mogące mieć zastosowanie w oszacowaniu parametrów modelu aproksymującego rozkłady płac i dochodów w Polsce.

2. Szacowanie parametrów rozkładu dochodów metodą największej wiarygodności

W analizie rozkładów płac i dochodów najczęściej rozważane są dwa rozkłady teoretyczne – rozkład logarytmiczno-normalny oraz rozkład Daguma. Pierwszy z tych rozkładów charakteryzuje prostota – zależy bowiem tylko od dwóch parametrów, które mają jasną interpretację ekonomiczną. Funkcję gęstości rozkładu logarytmiczno-normalnego wyraża następujący wzór:

$$f(y) = \frac{1}{y\sigma\sqrt{2\pi}} e^{-\left(\frac{\ln y - \mu}{\sigma}\right)^2}, \quad y \geq 0, \quad (1)$$

gdzie: μ – wartość oczekiwana logarytmów zmiennej losowej Y ,

σ – odchylenie standardowe logarytmów zmiennej losowej Y .

Wadą rozkładu logarymiczno-normalnego jest stosunkowo słabe dopasowanie do rozkładów empirycznych, szczególnie dla wysokich grup dochodowych.

Rozkład Daguma ma szereg własności (por. [2]), które pozwalają na jego lepsze dopasowanie do rozkładów płac i dochodów w różnych przekrojach. Jego funkcja gęstości zależy jednak od trzech parametrów:

$$f(y) = \beta \lambda y^{-\delta-1} (1 + \lambda y^{-\delta})^{-\beta-1}, \quad y > 0. \quad (2)$$

Dystrybuanta tego rozkładu przybiera postać:

$$F(y) = (1 + \lambda y^{-\delta})^{-\beta}, \quad y > 0. \quad (3)$$

Do szacowania parametrów modeli rozkładu dochodów najczęściej wykorzystywana jest metoda największej wiarygodności. W przypadku rozkładów dwuparametrowych (gamma, logarymiczno-normalny) estymatory NW mogą być podane w sposób bezpośredni.

Dla rozkładu logarymiczno-normalnego estymatory największej wiarygodności parametrów μ i σ^2 otrzymujemy z następujących wzorów (por. [1]):

$$m = \frac{1}{n} \sum_{i=1}^n \ln y_i, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (\ln y_i - m)^2. \quad (4)$$

Oba estymatory wyrażone wzorami (4) są nieobciążone i najbardziej efektywne, a ich wariancje dla dużych prób wynoszą:

$$D^2(m) = \frac{\sigma^2}{n}, \quad D^2(s^2) = \frac{2\sigma^2}{n-1} \cong \frac{2\sigma^2}{n}. \quad (5)$$

W przypadku danych pogrupowanych konieczne są jednak modyfikacje wzorów (4) [5]:

$$m' = \sum_{i=1}^k g_i \ln \bar{y}_i, \quad s^2 = \sum_{i=1}^k g_i \ln^2 \bar{y}_i - m'^2, \quad (6)$$

gdzie: $\ln \bar{y} = \frac{\ln y_{i-1} + \ln y_i}{2}$,

y_{i-1} – dolna granica grupy i ,

y_i – górna granica grupy i ,

g_i – frakcja jednostek w i -tej grupie,

k – liczba grup.

Jeżeli skrajne grupy są nieograniczone i liczne, możemy popełnić duży błąd przy obliczaniu średnich grupowych, a co za tym idzie – otrzymane estymatory będą miały gorsze właściwości, niż to wynika z poprzednich rozważań.

W przypadku bardziej skomplikowanych rozkładów (trój- i czteroparametrowych) zastosowanie metody największej wiarygodności wymaga posłużenia się metodami numerycznymi, które pozwalają znaleźć maksimum funkcji wiarygodności postaci:

$$\ln L = \sum_{i=1}^k \gamma_i \ln[F(y_i) - F(y_{i-1})], \quad (7)$$

gdzie: γ_i – częstości empiryczne,

$F(y)$ – dystrybuanta rozkładu teoretycznego,

k – liczba klas.

Dla rozkładu Daguma pierwszego typu jest to funkcja zmiennych: λ, β, δ .

Jeżeli przez θ oznaczymy wektor parametrów rozkładu teoretycznego, przez $\hat{\theta}$ zaś wektor ich estymatorów, to można zapisać:

$$\hat{\theta} \underset{n \rightarrow \infty}{\sim} N(\theta, \Sigma^{-1}), \quad (8)$$

gdzie: Σ – macierz informacji Fischera, której elementy określone są następująco:

$$\sigma_{ij} = nE \left[\frac{d \ln(f(y))}{d\theta_i} \cdot \frac{d \ln f(y)}{d\theta_j} \right]. \quad (9)$$

Zaletą estymatorów NW jest także własność niezmienniczości [9]. Miary będące funkcjami estymatorów NW są również estymatorami największej wiarygodności. Ma to zastosowanie przy estymacji współczynników nierównomierności, które można przedstawić jako funkcje parametrów rozkładu teoretycznego.

Niech I oznacza miarę nierównomierności, $\hat{I}$ zaś jej estymator będący funkcją $\hat{\theta}$:

$$\hat{I} \underset{n \rightarrow \infty}{\sim} N(I, V(I)).$$

Wariancja tego estymatora określona jest następująco:

$$V(\hat{I}) = \left[\frac{dI(\theta)}{d\theta} \right]^T \sum^{-1} \left[\frac{dI(\theta)}{d\theta} \right]. \quad (10)$$

W praktyce z wykorzystaniem metody NW do szacowania parametrów rozkładu dochodów wiążą się następujące problemy:

1. Liczebność próby nie jest na tyle duża, aby spełnione były asymptotyczne własności estymatorów największej wiarygodności.

2. Dostępne dane nie odpowiadają w pełni przyjętemu modelowi ze względu na luki lub obserwacje odstające. Przyczyny mogą być następujące:

- błędy odpowiedzi związane ze sposobem zbierania informacji (np. nieporozumienia związane z okresem badania),
- błędy opracowania danych związane z usuwaniem dochodów ujemnych i zerowych, „ucinięcie dochodów” wysokich,
- błędy związane z grupowaniem danych polegające na przesuwaniu dochodów z grupy do grupy,
- charakterystyczne dla płac i dochodów występowanie obserwacji nietypowych.

Wpływ tego typu odchyień od przyjętego modelu na własności estymatorów NW był badany m.in. przez van Praaga i Hagenaarsa [7]. Okazało się, że omawiane estymatory są obciążone i nieefektywne mimo dużych prób. Ponadto pogorszenie własności estymatorów jest tym większe, im mniej skomplikowany model. W przypadku modeli trój- i czteroparametrycznych dodatkowe parametry kształtu pozwalają na lepsze dopasowanie „ogonów” do obserwacji ekstremalnych.

W pierwszym przypadku (zbyt mało liczne próby) można wykorzystywać metodę estymacji zaproponowaną przez Stoppa [6]. Metoda ta pozwala otrzymać efektywne estymatory parametrów modeli rozkładów dochodów, dla których jawne formy estymatorów NW lub MNK nie mogą być podane. Chodzi tu o takie rozkłady, jak: uogólniony beta, uogólniony gamma, Pareto pierwszego i drugiego rodzaju, Singha-Maddali, Daguma. Estymatory te wyprowadzone są na podstawie własności dystrybucyjności $F(y)$ oraz elastyczności $\varepsilon(y, F(y))$ przy wykorzystaniu kryterium najmniejszych kwadratów.

W drugim przypadku lepsze wyniki można uzyskać, stosując metody estymacji odpornej, które pozwalają wyeliminować wpływ zanieczyszczeń modelu przy niewielkiej stracie na efektywności.

3. Metoda OBRE (*optimal b-robust estimation*)

Estymatorami odpornymi będziemy nazywać estymatory, które są niewrażliwe na odchylenia od przyjętego modelu rozkładu dochodów.

Estymatory OBRE należą do klasy estymatorów M będącej uogólnieniem estymatorów NW. Estymatory M zostały zaproponowane przez Hubera [4]. Standaryzowany estymator OBRE jest określony jako rozwiązanie względem θ następującego równania (por. [3]):

$$\sum_{i=1}^n [s(y_i; \theta) - \mathbf{a}] w_c(y_i; \mathbf{a}, \mathbf{A}) = 0, \quad (11)$$

gdzie: $w_c(y_i; \mathbf{a}, \mathbf{A}) = \min \left\{ 1; \frac{c}{\|\mathbf{A}[s(y_i; \boldsymbol{\theta}) - \mathbf{a}]\|} \right\}$ – wagi funkcji Hubera,

$\mathbf{a}$ – wektor określony następująco:

$$\mathbf{a} = E[s(y; \boldsymbol{\theta})w_c(y; \mathbf{a}, \mathbf{A})]; \quad (12)$$

$\mathbf{A}$ – macierz o wymiarach $p \times p$, przy czym:

$$\mathbf{A}^{-1}\mathbf{A}^{-T} = E[[s(y; \boldsymbol{\theta}) - \mathbf{a}][s(y; \boldsymbol{\theta}) - \mathbf{a}]^T \cdot w_c(y; \mathbf{a}, \mathbf{A})^2]; \quad (13)$$

$s(y_i, \boldsymbol{\theta})$ – wektor pochodnych logarytmu f -gęstości rozkładu dochodów (tzw. funkcja wynikowych, *scores function*),

c – stała wpływająca na efektywność i odporność estymatora (czym większe c tym mniej odporna estymacja, ale tym bardziej efektywna).

- Przy $c \Rightarrow \infty$ otrzymujemy estymatory NW, które są najefektywniejsze ale nie są odporne.
- Przy $c = \sqrt{p}$, gdzie p oznacza liczbę parametrów modelu, otrzymujemy estymatory najbardziej odporne.
- Estymatory OBRE są optymalne w tym sensie, że są odporne przy stosunkowo małej stracie na efektywność (dla rozkładu Daguma, przyjmując $c = 2$, uzyskuje się efektywność 0,95).

Macierze $\mathbf{a}$ i $\mathbf{A}$ kontrolują efektywność i zgodność estymatorów OBRE.

4. Badanie odporności estymatorów NW dla rozkładu Daguma pierwszego typu

Próba zawierająca zanieczyszczenia nie będzie realizacją rzeczywistego rozkładu $F(y)$, ale następującej mieszaniny:

$$F_g(y) = (1 - \varepsilon)F(y) + \varepsilon G(y). \quad (14)$$

Jedną z metod oceny odporności estymatorów jest wykorzystanie funkcji wpływu (*influence function*). Funkcja ta dla estymatora T parametru θ określona jest następująco:

$$IF(y) = \frac{\partial T(F_g(y))}{\partial \varepsilon} \bigg|_{\varepsilon \rightarrow 0}. \quad (15)$$

Określa ona wpływ, jaki może mieć na oszacowanie nieskończenie małe zanieczyszczenie w punkcie y .

W przypadku estymatorów NW funkcja wpływu (IF) jest proporcjonalna do funkcji wynikowej ($s(y)$). Dla parametrów modelu Daguma funkcja wynikowa jest następująca:

$$\begin{aligned}
s_1(y) &= \frac{\partial \log f(y)}{\partial \beta} = -\tilde{\beta}(1, \beta) - \log(1 + \lambda y^{-\delta}), \\
s_2(y) &= \frac{\partial \log f(y)}{\partial \lambda} = -(1/\lambda) - (\beta + 1) \frac{y^{-\delta}}{1 + \lambda y^{-\delta}}, \\
s_3(y) &= \frac{\partial \log f(y)}{\partial \delta} = -(1/\delta) + (\beta + 1) \frac{y^{-\delta} \log(y)}{1 + \lambda y^{-\delta}} - \log(y),
\end{aligned} \tag{16}$$

gdzie: $\tilde{\beta}(1, \beta) = \frac{dB(1, \beta)}{d\beta} / B(1, \beta)$.

W celu oceny odporności estymatorów parametrów λ, β, δ wyznaczone zostały następujące granice:

$$\begin{aligned}
&\bullet \lim_{y \rightarrow 0} s_1(y) = -\infty, \quad \lim_{y \rightarrow \infty} s_1(y) = \text{const}, \\
&\bullet \lim_{y \rightarrow 0} s_2(y) = \text{const}, \quad \lim_{y \rightarrow \infty} s_2(y) = \text{const}, \\
&\bullet \lim_{y \rightarrow 0} s_3(y) = \infty, \quad \lim_{y \rightarrow \infty} s_3(y) = -\infty.
\end{aligned} \tag{17}$$

$$\sup_y \|s(y)\| = \infty.$$

Funkcja wynikowa dla modelu Daguma jest więc nieograniczona. Oznacza to, że nawet pojedyncza obserwacja odstająca jest w stanie silnie wpłynąć na wartości estymatorów NW parametrów tego rozkładu, a co za tym idzie – doprowadzić do błędnych wniosków.

5. Algorytm szacowania parametrów rozkładu Daguma metodą OBRE

Etap I. Przyjmujemy wartości startowe parametrów $\theta = (\beta, \lambda, \delta)^T$ odpowiadające wartościom otrzymanym za pomocą MNW.

Etap II. Przyjmujemy wartości startowe wektora $\mathbf{a}$ i macierzy $\mathbf{A}$, które stanowią rozwiązanie równań dla $c = \infty$, czyli dla MNW.

$$\mathbf{a} = 0,$$

$$\mathbf{A} = \left[\sum^{1/2}(\theta) \right]^{-T} \tag{18}$$

gdzie: $\sum(\theta)$ – macierz informacji Fishera, której elementy σ_{ij} dla rozkładu Daguma są następujące:

$$\sigma_{11} = n / \beta^2,$$

$$\begin{aligned}
\sigma_{21} &= \frac{n}{\lambda(\beta+1)}, \\
\sigma_{22} &= \frac{n\beta}{\lambda^2(\beta+2)}, \\
\sigma_{31} &= \frac{n}{\delta(\beta+1)}[\psi(2) - \psi(\beta) - \ln \lambda], \\
\sigma_{32} &= \frac{n\beta}{\lambda\delta(\beta+2)}[\psi(2) - \Psi(\beta+1) - \ln \lambda], \\
\sigma_{33} &= \frac{n}{\delta^2} + \frac{n\beta}{(\beta+2)\delta^2} \left\{ [\psi(2) - \Psi(\beta+1) - \ln \lambda]^2 + \psi'(\beta+1) + \psi'(2) \right\},
\end{aligned} \tag{19}$$

gdzie: $\psi(\cdot)$ – funkcja digamma,
 β, λ, δ – parametry rozkładu Daguma.

Etap III. Ustalamy nowe wartości $\mathbf{a}$ i $\mathbf{A}$ na podstawie następujących wzorów:

$$\begin{aligned}
\mathbf{a} &= \int s(y, \boldsymbol{\theta}) w(y, \boldsymbol{\theta}) dF(y) / \int s(y, \boldsymbol{\theta}) dF(y), \\
\mathbf{A} &= [\mathbf{M}_2^{1/2}(\boldsymbol{\theta})]^{-T},
\end{aligned} \tag{20}$$

$$\text{gdzie: } \mathbf{M}_k(\boldsymbol{\theta}) = \int [s(y, \boldsymbol{\theta}) - \mathbf{a}][s(y, \boldsymbol{\theta}) - \mathbf{a}] w^k(y, \boldsymbol{\theta}) dF(y), \quad k = 1, 2. \tag{21}$$

Etap IV. Korzystając z metody Newtona-Raphsona, obliczamy $\Delta\boldsymbol{\theta}$

$$\Delta\boldsymbol{\theta} = \mathbf{M}_1^{-1}(\boldsymbol{\theta}) \left[\frac{1}{n} \sum_{i=1}^n (s(y, \boldsymbol{\theta}) - \mathbf{a}) w(y, \boldsymbol{\theta}) \right]^2, \tag{22}$$

gdy $|\Delta\boldsymbol{\theta}| > \varepsilon$ to wracamy do etapu 2, przyjmując za nowe parametry $\boldsymbol{\theta} = \boldsymbol{\theta} + \Delta\boldsymbol{\theta}$.

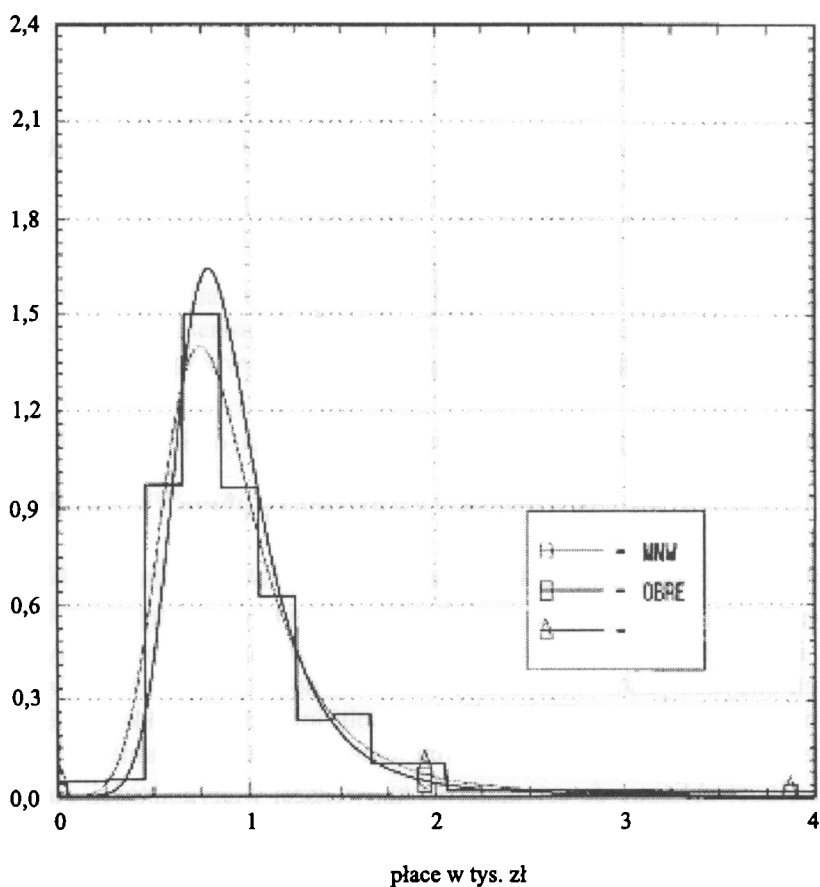
Wyniki obliczeń przedstawione zostały w tab. 1 oraz na rys. 1-3. Estymatory OBRE i największej wiarygodności szacowane były na podstawie danych indywidualnych otrzymanych z badania aktywności ekonomicznej ludności. Dane te dotyczyły rozkładu płac w województwie wielkopolskim w czwartym kwartale 2000 r.

Wraz ze wzrostem stałej c otrzymujemy estymatory parametrów λ, β i δ coraz bardziej zbliżone do estymatorów największej wiarygodności. Już dla $c = 5$ różnice między oszacowaniami otrzymanymi metodami OBRE i MNW są minimalne i pojawiają się dopiero na piątym miejscu po przecinku. Wynika to ze stosunkowo niewielkiego udziału zanieczyszczeń w rozważanym rozkładzie. Na rysunkach można zauważyć, że funkcje gęstości oszacowane metodą OBRE tym lepiej dopasowują się do środkowej części rozkładu, im mniejsza jest wartość stałej c (wzrost odporności). Co bardzo ważne, towarzyszy temu jedynie niewielki spadek efektywności estymatorów w porównaniu z MNW.

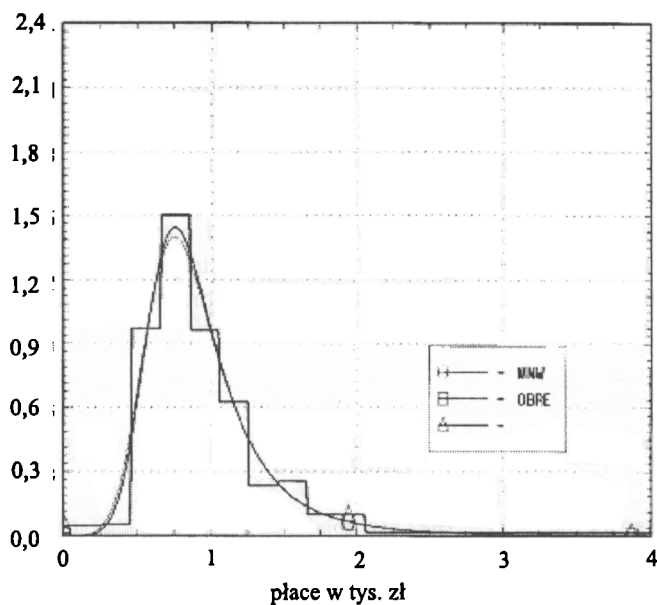
Tabela 1. Parametry rozkładu Daguma oszacowane metodą OBRE (na podstawie BAEL)

Stała c	Parametry rozkładu Daguma		
	$\hat{\lambda}$	$\hat{\beta}$	$\hat{\delta}$
$c = 10$	0,3260	1,4588	3,9766
$c = 5$	0,3260	1,4588	3,9766
$c = 4$	0,3106	1,5245	3,9469
$c = 3$	0,2884	1,6241	4,0020
$c = 2$	0,2867	1,5877	4,7625

Źródło: opracowanie własne.

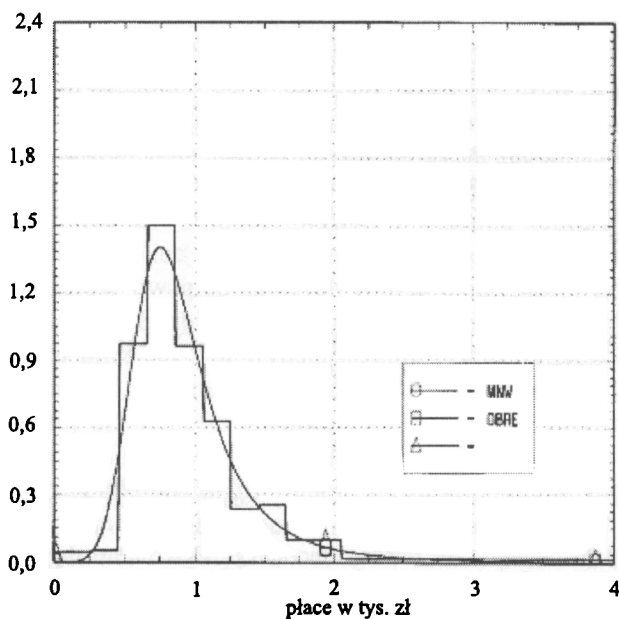
Rys. 1. Funkcje gęstości rozkładu Daguma dla płac w województwie wielkopolskim w listopadzie 2000 r. [$c = 2$]

Źródło: opracowanie własne.



Rys. 2. Funkcje gęstości rozkładu Daguma dla płac w województwie wielkopolskim w listopadzie 2000 r. [c = 3]

Źródło: opracowanie własne.



Rys. 3. Funkcje gęstości rozkładu Daguma dla płac w województwie wielkopolskim w listopadzie 2000 r. [c = 4]

Źródło: opracowanie własne.

6. Zakończenie

Szacowanie parametrów rozkładu płac i dochodów metodą OBRE może stanowić uzupełnienie metod klasycznych, do jakich należy metoda największej wiarygodności. Metoda ta pozwala w obiektywny sposób potraktować zbiór danych bez konieczności odrzucania informacji odstających na podstawie subiektywnych sądów badacza. Oprócz odpornych metod estymacji interesujące może być także rozważanie pewnych modyfikacji testów statystycznych stosowanych do weryfikacji hipotez o zgodności rozkładów – na przykład opartego na ilorazie wiarygodności testu Coksa.

Literatura

- [1] Aitchison I., Brown I.A.C. (1957), *The Lognormal Distribution*, Cambridge.
- [2] Dagum C. (1977), *A New Model of Personal Income Distribution. Specification and Estimation*, „Economie Appliquee”, XXX, s. 413-436.
- [3] Hampel F.R. et al. (1986), *The Approach Based on Influence Functions*, [w:] *Robust Statistics*, John Wiley, New York.
- [4] Huber P.I. (1964), *Robust Estimation of a Location Parameter*, „Annals of Mathematical Statistics” 35, s. 73-101.
- [5] Kordos J. (1973), *Metody analizy i prognozowania rozkładów płac i dochodów ludności*, Warszawa.
- [6] Stoppa G. (1995), *Explicit Estimators for Income Distributions*, [w:] „Research on Economic Inequality”, Jai Press, Greenwich vol. 6, s. 393-405.
- [7] Van Praag B.M.S., Hagenaars A.J.M. (1983), *The Influence of Classification and Observations Errors on the Measurement of Income Inequality*, „Econometrica” 51, nr 4, s. 1093-1108.
- [8] Victoria-Feser M.P. (2000), *Robust Methods for the Analysis of Income Distribution, Inequality and Poverty*, „International Statistical Review” 68, s. 277-293.
- [9] Zehna P. (1966), *Invariance of Maximum Likelihood Estimation*, „Annals of Mathematical Statistics” 37, s. 744-745.

REMARKS ON THE ESTIMATION OF INCOME DISTRIBUTION PARAMETERS

Summary

The parameters of income distribution models are usually estimated by means of the maximum likelihood method. The maximum likelihood estimators are known to have very good asymptotic properties – they are the most effective, among other things. Nevertheless they are not robust to any deviations from the assumed model. Statistical data concerning wage and income distributions by size are usually contaminated because of the method of gathering information. So the methods of robust estimation, which would find compromise between efficiency and robustness, could be very useful. In the paper we present the results of the estimation of wage distributions in Poland by means of the maximum likelihood method and the optimal B-robust method (OBRE). The basis for the calculation was the individual data coming from the Labour Force Survey.