Anna Sączewska-Piotrowska*

Identyfikacja determinant bogactwa dochodowego z zastosowaniem modelu logitowego

Wstęp

Przeprowadzane badania rozkładu dochodów dotyczą w głównej mierze nierówności oraz skupiają się na gospodarstwach domowych dysponujących najniższymi dochodami, czyli na gospodarstwach ubogich. Przeciwstawnym do ubóstwa dochodowego jest bogactwo dochodowe, które nie jest dobrze rozpoznanym zjawiskiem. Dotychczasowe badania bogactwa dochodowego dotyczą głównie jego zasięgu. Należy podkreślić, że zasięg bogactwa w różnych grupach gospodarstw domowych nie jest taki sam, ponieważ różne czynniki zwiększają lub zmniejszają szanse na wystąpienie tego zjawiska. Celem niniejszego opracowania jest identyfikacja czynników wpływających na bogactwo dochodowe gospodarstw domowych, a narzędziem umożliwiającym tę identyfikację jest model logitowy (model regresji logistycznej).

1. Bogactwo dochodowe – podstawowe pojęcia

Problemem pojawiającym się na początku badania bogactwa jest zdefiniowanie tego zjawiska. Bogactwo może być rozumiane jako stan posiadania odpowiadający wąskiej elicie majątkowej społeczeństwa, szczytom jego najzamożniejszych warstw [Żarnowski, 1992]. Bogactwo jest więc identyfikowane z najwyższym poziomem zamożności, przy czym poziom dochodów nie jest tożsamy z poziomem zamożności [Radziukiewicz, 2006, s. 12]. Bogactwo dochodowe jest pojęciem węższym niż bogactwo, ponieważ jest ono postrzegane jedynie przez pryzmat dochodów, będąc tym samym przeciwstawnym do ubóstwa dochodowego.

Po zdefiniowaniu bogactwa dochodowego należy przejść do wyznaczenia granicy bogactwa, czyli odpowiedzieć na pytanie o minimalną wysokość dochodów, jakie należy osiągnąć, aby zostać uznanym za bogatego. W badaniach empirycznych zamożności można spotkać granice opierające na bezwzględnej wielkości dochodów przypadających na

* Dr, Katedra Metod Statystyczno-Matematycznych w Ekonomii, Wydział Ekonomii, Uniwersytet Ekonomiczny w Katowicach, ul. 1 Maja 50, 40-287 Katowice, anna.saczewska-piotrowska@ue.katowice.pl

osobę lub na gospodarstwo domowe. Przykładowo, T. Słaby [*Konsumpcja*…, 2006, s. 8] terminem „elita ekonomiczna" określa grupę wysokodochodowych gospodarstw domowych, których dochody wynoszą powyżej 5000 zł miesięcznie na osobę. W badaniu przeprowadzonym przez KPMG przyjęto, że „osoby bogate i zamożne" to osoby osiągające miesięcznie dochód powyżej 7100 zł brutto [KPMG w Polsce, 2014]. Podejście takie ma niewątpliwie jedną wadę – tak wyznaczona granica musi być ciągle korygowana, ponieważ należy każdorazowo przy jej wyznaczaniu uwzględniać poziom inflacji.

Kolejna metoda wyznaczania granicy bogactwa dochodowego bazuje na udziale dochodu najbogatszych p% gospodarstw domowych. W badaniach empirycznych najczęściej przyjmowane jest 5% lub 1%, np. [*Top*…, 2007; Leigh, 2009]. Granicy bogactwa tak rozumianej nie wybrano w analizie w sposób celowy, ponieważ nie można wtedy analizować zmian odsetka bogatych gospodarstw, gdyż w każdym okresie odsetek ten jest równy przyjętemu poziomowi p%.

Granica bogactwa dochodowego może być również ustalana jako k-krotność mediany rozkładu dochodów ekwiwalentnych, przy czym przyjmuje się najczęściej dwu-, trzy- i czterokrotność mediany. W przeprowadzonej analizie przyjęto granicę bogactwa obliczoną jako dwukrotność mediany rozkładów dochodów ekwiwalentnych. Przyjęcie granicy bogactwa wyższej niż 200% mediany powoduje, że odsetek bogatych gospodarstw domowych jest bardzo niski, co uniemożliwia przeprowadzenie wiarygodnej analizy w grupach gospodarstw domowych ze względu na małe liczebności tych grup lub wręcz brak gospodarstw domowych w niektórych z wyróżnionych grup.

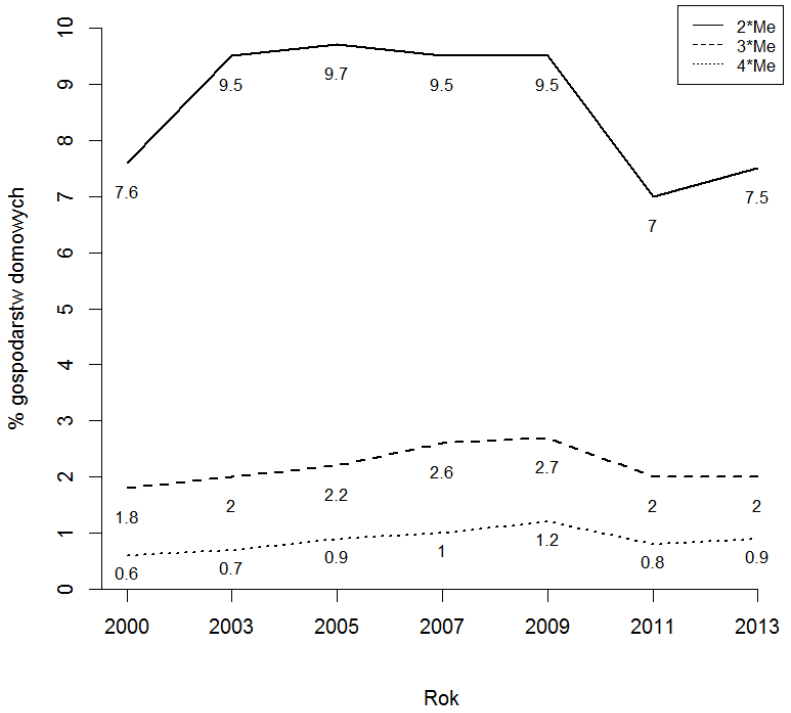
Mając ustaloną granicę bogactwa, można obliczyć wskaźniki statystyczne, pozwalające na analizę tego zjawiska. Podstawowym miernikiem jest stopa bogactwa dochodowego (*richness headcount ratio*), określona wzorem [Peichl i inni, 2008]:

$$R^{HC}(x,\rho)=\frac{1}{n}\sum_{i=1}^{n} I\left(x_i>\rho\right)=\frac{r}{n}, \tag{1}$$

gdzie: ρ – linia bogactwa, $I(\cdot)$ – funkcja wskaźnikowa, przyjmująca wartość 1, gdy gospodarstwo domowe jest bogate oraz 0 w przeciwnym wypadku, r – liczba bogatych gospodarstw domowych, n – liczba gospodarstw z dochodami $x_1, x_2, ..., x_n$. Stopa bogactwa informuje o udziale bogatych gospodarstw domowych w grupie gospodarstw ogółem.

Na podstawie danych projektu „Diagnoza społeczna" obliczono zasięg bogactwa dochodowego, wykorzystując w tym celu wspomnianą granicę bogactwa – dwukrotność mediany rozkładu dochodów ekwiwalentnych[1]. Zasięg bogactwa obliczono dla lat 2000, 2003, 2005, 2007, 2009, 2011 oraz 2013 (są to wszystkie lata, w których realizowano badanie). Granica bogactwa była obliczana osobno w każdym z badanych lat. Wyniki obliczeń zaprezentowano na rysunku 1.

Rysunek 1. Zasięg bogactwa dochodowego w Polsce w latach 2000–2013



Źródło: Opracowanie własne na podstawie [Rada Monitoringu Społecznego, 2014].

Stosując jako granicę bogactwa dochodowego trzy- i czterokrotność mediany, odsetki bogatych gospodarstw domowych w latach 2000–2013 były dosyć niskie – nie przekroczyły odpowiednio 2,7% oraz 1,2%. Przyjmując jako linię 200% mediany rozkładu dochodów ekwiwalent-

[1] Wszystkie obliczenia i wykresy wykonano w programie R [R Development Core Team, 2015].

nych, odsetki bogatych gospodarstw wahały się w granicach od 7,0% do 9,7%. Spadek udziału bogatych gospodarstw domowych (według wszystkich trzech granic bogactwa) miał miejsce w 2011 r., u podłoża czego stał niewątpliwie kryzys finansowy, który wpłynął negatywnie na budżety gospodarstw domowych. W kolejnym kroku zbadano, które czynniki wpływają na bogactwo dochodowe gospodarstw domowych, stosując w tym celu model logitowy.

2. Dwumianowy model logitowy

Model logitowy może być modelem dwumianowym lub wielomianowym. Dwumianowy model logitowy jest modelem, którego można użyć w celu opisania wpływu zmiennych $X_1, X_2, ..., X_k$ (jakościowych lub ilościowych) na dychotomiczną zmienną Y. W przypadku modelu wielomianowego zmienna objaśniana Y przyjmuje więcej niż dwie wartości. W przeprowadzonej analizie zmienna objaśniana przyjmowała dwie wartości, stąd właściwą postacią był model dwumianowy.

Niech Y oznacza zmienną dychotomiczną o wartościach: 1 – jeżeli dany wariant wystąpi, 0 – jeżeli dany wariant nie wystąpi. Wówczas [Stanisz, 2007, s. 219–220]:

$$p = P(Y = 1 | X_1 = x_1, X_2 = x_2, ..., X_k = x_k) = \frac{\exp\left(a_0 + \sum_{i=1}^{k} a_i x_i\right)}{1 + \exp\left(a_0 + \sum_{i=1}^{k} a_i x_i\right)}, \qquad (2)$$

gdzie $a_i (i = 0,1,...,k)$ są współczynnikami regresji. Model (2) jest więc modelem wiążącym prawdopodobieństwo jednego z dwóch możliwych wyników zmiennej Y ze zmiennymi objaśniającymi. Współczynniki regresji są zazwyczaj estymowane metodą największej wiarygodności. Wartości $\exp(a_i)$ w modelu (2) są najczęściej interpretowane za pomocą pojęcia ilorazu szans (*odds ratio*). Szansa jest definiowana jako prawdopodobieństwo wystąpienia zdarzenia do prawdopodobieństwa niewystąpienia zdarzenia. Iloraz szans dwóch porównywanych grup A i B definiowany jest jako stosunek „szansy" wystąpienia A do „szansy" wystąpienia B. W modelu logitowym w przypadku zmiennej dychotomicznej X_i iloraz szans pokazuje, ilokrotnie zmienia się szansa u jednostki, dla której $X_i = 1$ względem jednostki, dla której $X_i = 0$, przy niezmienionych wartościach pozostałych zmiennych objaśniających. Wyrażenie $\exp(a_0)$ jest równe szansie dla grupy referencyjnej, tzn. gru-

py, w której wszystkie zmienne objaśniające są równe zero. Gdy zmienna X_i jest zmienną ilościową, to iloraz szans mówi, jak zmieni się szansa, jeżeli zmienna X_i wzrośnie o jedną jednostkę przy pozostałych zmiennych ustalonych [Jackowska, 2011].

Do testowania statystycznej istotności poszczególnych parametrów modelu można zastosować standardowy test *t* Studenta, natomiast do testowania statystycznej istotności wszystkich parametrów przy zmiennych objaśniających – test ilorazu wiarygodności (tzw. *LR* test). W teście *LR* hipoteza zerowa głosi, że wszystkie parametry są równe zero, natomiast hipoteza alternatywna, że przynajmniej jeden z parametrów jest różny od zera. Statystyka ilorazu wiarygodności jest określona wzorem [Gruszczyński, 2001, s. 64; Książek, 2013, s. 60–61]:

$$LR = -2\left(\ln L_0 - \ln L_{FM}\right), \tag{3}$$

gdzie L_{FM} jest maksymalną wiarygodnością oszacowanego modelu (zawierającego zmienne objaśniające), L_0 jest maksymalną wiarygodnością modelu ograniczonego, zawierającego jedynie wyraz wolny. Statystyka *LR* ma dla dużych prób rozkład χ^2 z *k* stopniami swobody, gdzie *k* jest liczbą zmiennych objaśniających w modelu.

Jakość zbudowanego modelu można również ocenić, korzystając z testu Hosmera-Lemeshowa, który dla różnych podgrup danych (najczęściej dla grup decylowych) porównuje obserwowane liczebności i oczekiwane liczebności występowania wartości wyróżnionej. Hipoteza zerowa głosi, że obserwowane i oczekiwane liczebności są równe we wszystkich wyróżnionych podgrupach, natomiast hipoteza alternatywna, że różnią się one w przynajmniej jednej podgrupie. Statystyka testowa ma postać [Więckowska, 2015, s. 319]:

$$HL = \sum_{g=1}^{G} \frac{\left(O_g - E_g\right)^2}{E_g\left(1 - \dfrac{E_g}{N_g}\right)}, \tag{4}$$

gdzie O_g to obserwowane liczebności, E_g to oczekiwane liczebności, N_g to liczba obserwacji w grupie *g*, *G* to liczba podgrup. Statystyka ta ma asymptotycznie (dla dużych liczności) rozkład χ^2 z $G-2$ stopniami swobody. Należy podkreślić, że w przypadku Hosmera-Lemeshowa brak podstaw do odrzucenia hipotezy zerowej jest pożądany, ponieważ wskazuje na podobieństwo liczności obserwowanych i oczekiwanych.

Miarą dopasowania modelu jest również miara zaproponowana przez McFaddena (tzw. pseudo-R^2) określona wzorem [Stanisz, 2007, s. 251]:

$$R^2_{McFadden} = 1 - \frac{\ln L_{FM}}{\ln L_0}. \tag{5}$$

Pseudo-R^2 bazuje na porównaniu wartości funkcji wiarygodności w oszacowanym modelu i modelu bez zmiennych objaśniających. Miara ta przyjmuje wartości z zakresu [0,1], należy jednak podkreślić, że w modelach logitowych niska wartość pseudo-R^2, zwłaszcza przy dużych zbiorach danych, nie świadczy o złym dopasowaniu modelu [Gruszczyński, 2001, s. 56]. Jak podkreśla D. McFadden [1977], wartości z zakresu 0,2–0,4 świadczą o bardzo dobrym dopasowaniu modelu do danych.

Do oceny jakości dopasowania modelu można również zastosować kryterium informacyjne Akaikego *AIC*, które pozwala porównać ze sobą modele różniące się jedynie zestawem zmiennych objaśniających. Kryterium informacyjne Akaikego wyraża się wzorem [Książek, 2013, s. 61]:

$$AIC = -2 \ln L_{FM} + 2k. \tag{6}$$

Do opisu badanego zjawiska należy wybrać model o minimalnej wartości *AIC*.

Często najważniejszą miarą dopasowania w modelach logitowych jest ich zdolność predyktywna. Należy podkreślić, ze termin „prognoza" w odniesieniu do danych przekrojowych dotyczy pewnej jednostki obserwacji, a nie jednostki czasu. Mikroprognozy mogą dotyczyć jednostek znajdujących się w próbie, a także jednostek spoza próby. Model logitowy pozwala ustalić mikroprognozy: prognozę $\hat{p}_i$ prawdopodobieństwa p_i oraz prognozę $\hat{y}_i$ wartości y_i (1 lub 0), tzn. mikroprognozę zmiennej Y dla *i*-tej jednostki obserwacji [Gruszczyński, 2001, s. 78].

Prognozę $\hat{p}_i$ wyznacza się jednoznacznie, pod warunkiem dysponowania danymi liczbowymi o zmiennych objaśniających. Wartości teoretyczne zmiennej objaśnianej $\hat{y}_i$ można wyznaczyć według standardowej zasady prognozy:

$$\hat{y}_i = \begin{cases} 1 \text{ dla } \hat{p}_i > 0{,}5 \\ 0 \text{ dla } \hat{p}_i \leq 0{,}5 \end{cases}. \tag{7}$$

W próbach niezbilansowanych (liczba wartości $y_i = 1$ znacznie różni się od liczby wartości $y_i = 0$) do prognozowania wartości teoretycznych powinno się przyjąć zasadę [Gruszczyński, 2001, s. 80]:

$$\hat{y}_i = \begin{cases} 1 \text{ dla } \hat{p}_i > p^* \\ 0 \text{ dla } \hat{p}_i \leq p^* \end{cases}, \tag{8}$$

gdzie p^* jest nową wartością odcinającą (*cut-off point*), wyznaczoną dla danej próby oraz dla danego badania. Dla wybranego punktu odcięcia można zbudować tablicę trafności oraz obliczyć na jej podstawie następujące mierniki [Gruszczyński, 2001, s. 83–84; Dudek, Dybciak, 2006; Jackowska, Wycinka, 2009; Harańczyk, 2010]:

1. Skuteczność reguły decyzyjnej (*accuracy*), zwana również zliczeniowym R^2, określająca udział poprawnie prognozowanych przez model przypadków w łącznej liczbie przypadków:

$$ACC = \frac{n_{00} + n_{11}}{n}, \tag{9}$$

gdzie n_{00} jest liczbą obserwacji, dla których $y_i = \hat{y}_i = 0$, natomiast n_{11} jest liczbą obserwacji, dla których $y_i = \hat{y}_i = 1$, n to liczba obserwacji.

2. Czułość (*sensitivity*) będąca proporcją obserwacji trafnie przewidywanych przez model „jedynek" w ogólnej liczbie zaobserwowanych „jedynek":

$$SENS = \frac{n_{11}}{n_{1\bullet}}, \tag{10}$$

gdzie $n_{1\bullet}$ jest sumą $y_i = 1$, niezależnie od tego, czy $\hat{y}_i = 1$ czy $\hat{y}_i = 0$.

3. Specyficzność (*specifity*) określająca udział trafnie przewidzianych przez model „zer" w grupie zaobserwowanych „zer":

$$SPEC = \frac{n_{00}}{n_{0\bullet}}, \tag{11}$$

gdzie $n_{0\bullet}$ jest sumą $y_i = 0$ niezależnie od tego, czy $\hat{y}_i = 1$ czy $\hat{y}_i = 0$.

Jeżeli mamy do czynienia z milionem obserwacji, to istnieje milion potencjalnych punktów odcięcia, czyli milion tablic trafności do przeanalizowania, spośród których należy wybrać tę z najlepszym podziałem. Aby dokonać tego wyboru, warto wykorzystać krzywe *ROC* (*receiver operating characteristic*), nie tylko po to, aby znaleźć optymalny punkt, ale również ocenić jakość skonstruowanego modelu. Konstrukcja krzywej

ROC wygląda następująco: dla każdego z punktów odcięcia należy obliczyć czułość i specyficzność, a następnie zaznaczyć otrzymane wyniki na wykresie. Tradycyjnie zaznacza się je w układzie współrzędnych, gdzie na osi odciętych jest „1-specyficzność", a na osi rzędnych „czułość". Uzyskane punkty należy ze sobą połączyć. Im więcej różnych wartości badanego wskaźnika, tym gładsza uzyskana krzywa. Jeśli przyjmujemy równe koszty błędnych klasyfikacji, to optymalnym punktem odcięcia jest punkt krzywej *ROC* znajdujący się najbliżej punktu o współrzędnych (0,1) [Harańczyk, 2010]. Drugim, często stosowanym w praktyce prostym kryterium wyboru punktu odcięcia jest przyjęcie udziału jedynek w próbie [Jackowska, Wycinka, 2009].

W celu oceny jakości modelu na podstawie krzywej *ROC* można wyliczyć pole pod wykresem krzywej, oznaczane jako *AUC* (*area under curve*), i traktować go jako miarę dobroci i trafności danego modelu. Jakość klasyfikacyjna modelu jest dobra, gdy krzywa znajduje się powyżej przekątnej $y = x$, czyli gdy *AUC* jest większe od 0,5. W tym celu testuje się hipotezę zerową mówiącą o tym, że pole pod wykresem krzywej *ROC* jest równe 0,5 (czyli wartości minimalnej). Statystyka testowa ma postać [Więckowska, 2015, s. 319]:

$$Z = \frac{A\hat{U}C - 0{,}5}{\sqrt{V\hat{a}r(A\hat{U}C)}} \tag{12}$$

gdzie $V\hat{a}r(A\hat{U}C)$ jest estymatorem wariancji pola $A\hat{U}C$. Statystyka Z ma asymptotycznie (dla dużych liczności) rozkład normalny. Nieodrzucenie hipotezy zerowej oznacza, że model nie ma żadnej mocy predykcyjnej [Kopczewska i inni, 2009, s. 532–533].

3. Determinanty bogactwa dochodowego w Polsce

Analizę determinant bogactwa dochodowego przeprowadzono dla 2013 r. z wykorzystaniem danych projektu „Diagnoza społeczna". W badaniu wzięło udział prawie 11 tys. gospodarstw domowych. Jako granicę bogactwa przyjęto dwukrotność mediany rozkładu dochodów ekwiwalentnych. Zmienną zależną w modelu logitowym była zmienna zerojedynkowa:

$$Y = \begin{cases} 1, \text{ gdy gospodarstwo domowe jest bogate,} \\ 0, \text{ gdy gospodarstwo domowe nie jest bogate,} \end{cases} \tag{13}$$

Zmienne niezależne były zmiennymi jakościowymi, które przedstawiono w postaci układów zmiennych zero-jedynkowych w taki sposób, że zmienna mająca m wariantów jest reprezentowana przez m–1 zmiennych zero-jedynkowych (w ten sposób uniknięto zjawiska współliniowości). W modelu uwzględnione zostały zmienne dotyczące płci, wieku i wykształcenia głowy gospodarstwa domowego, klasy miejscowości zamieszkania, liczby osób w gospodarstwie, grupy społeczno-ekonomicznej, statusu gospodarstwa na rynku pracy, obecności dzieci do lat 14 oraz województwa zamieszkiwanego przez gospodarstwo. Model logitowy szacowano w dwóch wersjach: w modelu 1 uwzględniono wszystkie wymienione zmienne, natomiast w modelu 2 usunięto zmienną, której wszystkie kategorie były statystycznie nieistotne. Próg statystycznej istotności ustalono na poziomie 0,1. Wyniki estymacji modelu 1 przedstawiono w tablicy 1.

Tablica 1. Wyniki estymacji modelu 1

Zmienne	Współczynnik	Iloraz szans
Stała	–2,497***	x
Płeć głowy gospodarstwa domowego:		
mężczyzna	ref.	
kobieta	–0,871***	0,418
Wiek głowy gospodarstwa domowego:		
25–34 lata	ref.	
35–44 lata	0,297*	1,346
45–59 lat	0,491***	1,635
60 i więcej lat	0,454*	1,574
Wykształcenie głowy gospodarstwa domowego:		
podstawowe i niższe	ref.	
zasadnicze zawodowe/gimnazjum	–0,161	0,851
średnie	0,951***	2,589
podyplomowe i wyższe	2,303***	10,005
Klasa miejscowości zamieszkania:		
miasta powyżej 500 tys.	ref.	
miasta 200–500 tys.	–0,165	0,848
miasta 100–200 tys.	–0,602**	0,547
miasta 20–100 tys.	–0,629***	0,533
miasta poniżej 20 tys.	–0,619***	0,538
wieś	–0,725***	0,484

Zmienne	Współ-czynnik	Iloraz szans
Liczba osób w gospodarstwie domowym:		
1	ref.	
2	0,269*	1,309
3	–0,285.	0,752
4	–0,500**	0,607
5	–1,009***	0,365
6 i więcej	–1,147***	0,318
Grupa społeczno-ekonomiczna:		
pracownicy	ref.	
rolnicy	–0,487.	0,614
pracujący na własny rachunek	0,443**	1,557
emeryci i renciści	–1,261***	0,283
utrzymujący się z niezarobkowych źródeł	–2,226**	0,108
Status gospodarstwa na rynku pracy:		
przynajmniej jedna osoba bezrobotna	–1,476***	0,228
brak osób bezrobotnych	ref.	
Dzieci w gospodarstwie domowym:		
gospodarstwa z dziećmi do lat 14	0,075	1,078
gospodarstwa bez dzieci do lat 14	ref.	
Województwo:		
dolnośląskie	–0,526**	0,591
kujawsko-pomorskie	–0,811***	0,444
lubelskie	–0,832***	0,435
lubuskie	–0,333	0,717
łódzkie	–0,868***	0,420
małopolskie	–1,021***	0,360
mazowieckie	ref.	
opolskie	–0,220	0,803
podkarpackie	–1,197***	0,302
podlaskie	–0,722**	0,486
pomorskie	–0,207	0,813
śląskie	–0,418*	0,659
świętokrzyskie	–1,129***	0,324
warmińsko-mazurskie	–0,682**	0,506
wielkopolskie	–0,758***	0,469
zachodniopomorskie	–0,515*	0,598

$.p < 0{,}1;{*}p < 0{,}05;{**}\ p < 0{,}01;{***}p < 0{,}001$

Źródło: Opracowanie własne na podstawie [Rada Monitoringu Społecznego, 2014].

Jak można zauważyć, nie wszystkie zmienne uwzględnione w modelu 1 okazały się istotne statystycznie. Na szanse pobytu w sferze bogactwa nie ma wpływu zasadnicze wykształcenie głowy gospodarstw domowego (w porównaniu do wykształcenia podstawowego), miejsce zamieszkania w miastach 200–500 tys. (w porównaniu do miast 500 tys.) i województwach: lubuskim, opolskim i pomorskim (w porównaniu do województwa mazowieckiego) oraz pobyt w gospodarstwie domowym dzieci do lat 14 (w porównaniu do braku dzieci w tym wieku). Zmienną odnoszącą się do obecności dzieci usunięto z modelu i w ten sposób uzyskano postać modelu 2 (tablica 2). Modele 1 i 2 poddano weryfikacji, której wyniki zawarto w tablicy 3.

Tablica 2. Wyniki estymacji modelu 2

Zmienne	Współczynnik	Iloraz szans
Stała	–1,976***	x
Płeć głowy gospodarstwa domowego:		
mężczyzna	ref.	
kobieta	–0,871***	0,418
Wiek głowy gospodarstwa domowego:		
25–34 lata	ref.	
35–44 lata	0,303*	1,354
45–59 lat	0,462***	1,587
60 i więcej lat	0,426*	1,530
Wykształcenie głowy gospodarstwa domowego:	ref.	
podstawowe i niższe	–0,165	0,848
zasadnicze zawodowe/gimnazjum	0,950***	2,586
średnie	2,304***	10,011
podyplomowe i wyższe		
Klasa miejscowości zamieszkania:		
miasta powyżej 500 tys.	ref.	
miasta 200–500 tys.	–0,166	0,847
miasta 100–200 tys.	–0,603**	0,547
miasta 20–100 tys.	–0,627***	0,534
miasta poniżej 20 tys.	–0,619***	0,538
wieś	–0,725***	0,484

Zmienne	Współ-czynnik	Iloraz szans
Liczba osób w gospodarstwie domowym:		
1	ref.	
2	0,272*	1,313
3	–0,262.	0,770
4	–0,468**	0,627
5	–0,965***	0,381
6 i więcej	–1,095***	0,335
Grupa społeczno-ekonomiczna:		
pracownicy	ref.	
rolnicy	–0,488.	0,614
pracujący na własny rachunek	0,443**	1,557
emeryci i renciści	–1,257***	0,285
utrzymujący się z niezarobkowych źródeł	–2,222**	0,108
Status gospodarstwa na rynku pracy:		
przynajmniej jedna osoba bezrobotna	–1,478***	0,228
brak osób bezrobotnych	ref.	
Województwo:		
dolnośląskie	–0,523**	0,593
kujawsko-pomorskie	–0,808***	0,446
lubelskie	–0,828***	0,437
lubuskie	–0,329	0,720
łódzkie	–0,867***	0,420
małopolskie	–1,022***	0,360
mazowieckie	ref.	
opolskie	–0,220	0,802
podkarpackie	–1,198***	0,302
podlaskie	–0,722**	0,486
pomorskie	–0,207	0,813
śląskie	–0,416*	0,660
świętokrzyskie	–1,129***	0,324
warmińsko-mazurskie	–0,687**	0,503
wielkopolskie	–0,757***	0,469
zachodniopomorskie	–0,513*	0,599

$.p < 0{,}1; * p < 0{,}05; ** p < 0{,}01; *** p < 0{,}001$

Źródło: Opracowanie własne na podstawie [Rada Monitoringu Społecznego, 2014].

Tablica 3. Zestawienie wyników weryfikacji oszacowanych modeli logitowych

Wyszczególnienie	Model 1	Model 2
AIC	4445,1	4443,6
$R^2_{McFadden}$	0,251	0,251
LR test:		
liczba stopni swobody	38	37
χ^2	1465,579	1465,358
wartość *p*	0,000***	0,000***
Test Hosmera-Lemeshowa:		
liczba stopni swobody	8	8
χ^2	7,309	10,693
wartość p	0,504	0,220

$.p < 0{,}1; * p < 0{,}05; ** p < 0{,}01; *** p < 0{,}001$

Źródło: Opracowanie własne na podstawie [Rada Monitoringu Społecznego, 2014].

Kryterium informacyjne Akaikego wskazuje, że lepszym modelem jest model 2. Pseudo-R^2 informuje, że obydwa modele są bardzo dobrze dopasowane do danych, a wartości tego miernika są dla obydwu modeli praktycznie takie same. W przypadku modeli 1 i 2 test ilorazu wiarygodności wskazuje, że przynajmniej jeden z parametrów istotnie różni się od zera (wszystkie parametry łącznie są istotne statystycznie), natomiast na podstawie testu Hosmera-Lemeshowa można stwierdzić, że liczebności obserwowane i teoretyczne nie różnią się istotnie w grupach decylowych. W dalszej części badania skupiono się na modelu 2, zawierającym mniej zmiennych i jednocześnie wskazanym przez *AIC* jako lepszy model.

Analizując ilorazy szans w tablicy 2, można stwierdzić, że szansa pobytu gospodarstwa domowego w sferze bogactwa była:

- o 58% niższa w gospodarstwach domowych kobiet niż mężczyzn,
- prawie 2,5-krotnie wyższa w gospodarstwach, których głowa ma średnie wykształcenie i 10-krotnie wyższa w gospodarstwach z głową z wyższym/podyplomowym wykształceniem niż w gospodarstwach, których głowa ma co najwyżej wykształcenie podstawowe,
- wyższa w gospodarstwach domowych, których głowa ma co najmniej 35 lat niż w gospodarstwach, których głowa ma 25–34 lata,
- niższa w gospodarstwach co najmniej 3-osobowych oraz wyższa w gospodarstwach 2-osobowych w porównaniu do 1-osobowych,

- wyższa w gospodarstwach pracujących na własny rachunek oraz niższa w gospodarstwach z pozostałych grup społeczno-ekonomicznych w porównaniu do gospodarstw pracowników,
- niższa o 77% w gospodarstwach z przynajmniej jedną osobą bezrobotną niż w gospodarstwach bez osób bezrobotnych,
- niższa o ponad 70% w gospodarstwach zamieszkujących województwa: małopolskie, podkarpackie i świętokrzyskie w porównaniu do gospodarstw z województwa mazowieckiego.

Badana próba nie była zbilansowana – zdecydowanie więcej gospodarstw było niebogatych niż bogatych. Jako punkt odcięcia wybrano częstość występowania bogatych gospodarstw domowych, czyli 0,075. Dla takiego punktu obliczono liczbę poprawnie prognozowanych przypadków (tablica 4) oraz porównano uzyskane wyniki z poprawnie prognozowanymi przypadkami dla standardowego punktu odcięcia 0,5 (tablica 5).

Tablica 4. Klasyfikacja przypadków dla punktu odcięcia 0,075

Obserwowane wartości zmiennej objaśnianej	Przewidywane wartości zmiennej objaśnianej		
	$\hat{y}_i = 0$	$\hat{y}_i = 1$	Razem
$y_i = 0$	7839	2293	10 132
$y_i = 1$	173	648	821
Razem	8012	2941	10 953

Źródło: Opracowanie własne na podstawie [Rada Monitoringu Społecznego, 2014].

Tablica 5. Klasyfikacja przypadków dla punktu odcięcia 0,5

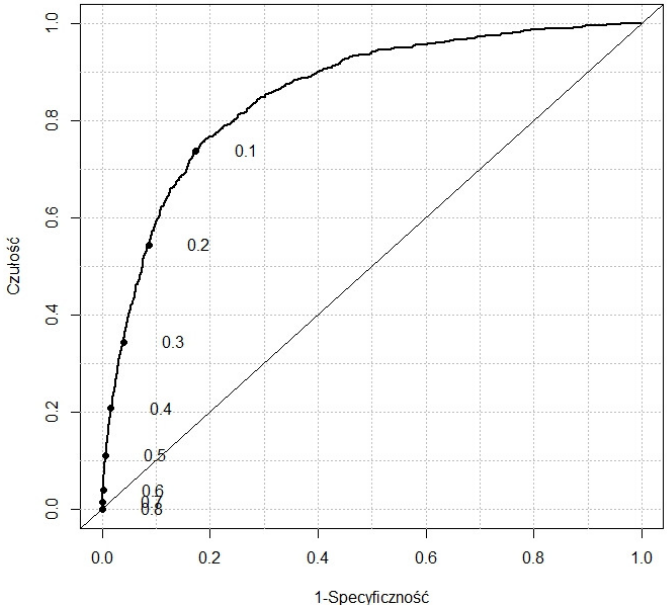
Obserwowane wartości zmiennej objaśnianej	Przewidywane wartości zmiennej objaśnianej		
	$\hat{y}_i = 0$	$\hat{y}_i = 1$	Razem
$y_i = 0$	10 069	63	10 132
$y_i = 1$	731	90	821
Razem	10 800	153	10 953

Źródło: Opracowanie własne na podstawie [Rada Monitoringu Społecznego, 2014].

Wykorzystując dane z tablicy 4, obliczono specyficzność, czułość oraz zliczeniowy R^2, które pozwoliły ocenić procentową trafność prognoz. Procent prawidłowych predykcji wyniósł 77,5%, przy czym specyficzność wyniosła 77,4%, a czułość 78,9%. Można na tej podstawie sądzić, że model przewiduje w nieco lepszym stopniu sukces (78,9% bogatych gospodarstw domowych zostało uznanych przez model jako bogate) niż porażkę (77,4% niebogatych gospodarstw zostało przewidziane przez model jako niebogate). W przypadku punktu odcięcia 0,5 zliczeniowy R^2 wyniósł 92,8%, specyficzność 99,4% oraz czułość 11%. Można więc zauważyć, że zliczeniowy R^2 jest zdecydowanie lepszy w przypadku wybrania jako punktu odcięcia wartości 0,5, lecz dzieje się to kosztem błędnego zaklasyfikowania gospodarstw bogatych (tylko co dziesiąte gospodarstwo bogate zostało uznane przez model jako bogate). Uznano więc, że klasyfikacja przypadków dla punktu odcięcia 0,075 jest lepsza mimo mniejszego odsetka poprawnie zaklasyfikowanych gospodarstw niebogatych.

Na podstawie wszystkich wartości czułości i specyficzności zbudowano krzywą *ROC* (rysunek 2).

Rysunek 2. Krzywa ROC



Źródło: Opracowanie własne na podstawie [Rada Monitoringu Społecznego, 2014].

Można zauważyć, że krzywa *ROC* jest wygięta w kierunku punktu o współrzędnych (0,1), a tym samym pole *AUC* jest dużo większe niż 0,5 i wynosi 0,856. Przeprowadzony test (na poziomie $p = 0{,}000$) potwierdza, że *AUC* jest istotnie większe niż 0,5, co oznacza, że jakość klasyfikacyjna modelu jest dobra i model może służyć do budowy prognoz. Na podstawie modelu zbudowano przykładowe prognozy pobytu w sferze bogactwa dochodowego gospodarstw domowych o różnych cechach:

– głowa gospodarstwa: mężczyzna, 40 lat z wykształceniem wyższym, 2-osobowe gospodarstwo pracowników bez osób bezrobotnych z dużego miasta (ponad 200 tys. mieszkańców) w województwie pomorskim: prognozowane prawdopodobieństwo wynosi $\hat{p}_1 = 0{,}627$, czyli na podstawie przyjętego punktu odcięcia $p^* = 0{,}075$ można się spodziewać, że gospodarstwo będzie bogate $(\hat{y}_1 = 1)$;
– głowa gospodarstwa domowego: kobieta, 30 lat z wykształceniem średnim, 3-osobowe gospodarstwo rolników z jedną osobą bezrobotną zamieszkujące wieś w województwie podlaskim: prawdopodobieństwo wynosi $\hat{p}_2 = 0{,}024$, czyli gospodarstwo nie będzie należeć do sfery bogactwa $(\hat{y}_2 = 0)$.

Zakończenie

Na podstawie oszacowanego modelu logitowego można stwierdzić, że szanse gospodarstwa domowego na pobyt w sferze bogactwa zależą w sposób istotny zarówno od cech samego gospodarstwa, jak i jego głowy. Szczególnie należy podkreślić wpływ wykształcenia głowy gospodarstwa domowego, grupy społeczno-ekonomicznej gospodarstwa oraz obecności osób bezrobotnych w gospodarstwie domowym.

W literaturze [np. Kasprzyk, Fura, 2011; Rusnak, 2012] można się spotkać z oszacowanymi modelami logitowymi ryzyka ubóstwa. Należy jednak zaznaczyć, że determinanty bogactwa i ubóstwa nie muszą się nawzajem uzupełniać, ponieważ pewne cechy mogą zwiększać szanse pobytu gospodarstwa w sferze bogactwa, ale nie muszą jednocześnie zmniejszać szans na pobyt w sferze ubóstwa. Nie można bowiem zapomnieć, że mogą istnieć grupy „średniaków", których rozkłady dochodów są dosyć równomierne i tym samym odsetek gospodarstw ubogich i bogatych w tych grupach jest niewielki. Przedmiotem kolejnych badań będzie porównanie determinant bogactwa i ubóstwa dochodowego, które pozwoli zweryfikować powyższą hipotezę.

Literatura

1. Dudek H., Dybciak M. (2006), *Zastosowanie modelu logitowego do analizy wyników egzaminu*, Zeszyty Naukowe SGGW, „Ekonomika i Organizacja Gospodarki Żywnościowej", nr 60.
2. Gruszczyński M. (2001), *Modele i prognozy zmiennych jakościowych w finansach i bankowości*, Szkoła Główna Handlowa, Warszawa.
3. Harańczyk G. (2010), *Krzywe ROC, czyli ocena jakości klasyfikatora i poszukiwanie optymalnego punktu odcięcia*, w: *Medycyna i analiza danych*, StatSoft, Kraków.
4. Jackowska B. (2011), *Efekty interakcji między zmiennymi objaśniającymi w modelu logitowym w analizie zróżnicowania ryzyka zgonu*, „Przegląd Statystyczny" nr 1–2.
5. Jackowska B., Wycinka E. (2009), *Modele ryzyka skreślenia z listy studentów na przykładzie studentów trybu niestacjonarnego*, w: „Taksonomia" nr 16. *Klasyfikacja i analiza danych – teoria i zastosowania*, Jajuga K., Walesiak M. (red.), Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 7, Wrocław.
6. Kasprzyk B., Fura B. (2011), *Wykorzystanie modeli logitowych do identyfikacji gospodarstw domowych zagrożonych ubóstwem*, „Wiadomości Statystyczne" nr 6.
7. *Konsumpcja elit ekonomicznych w Polsce – ujęcie empiryczne* (2006), Słaby T. (red.), SGH, Warszawa.
8. Kopczewska K., Kopczewski T., Wójcik P. (2009), *Metody ilościowe w R. Aplikacje ekonomiczne i finansowe*, CeDeWu, Warszawa.
9. KPMG w Polsce (2014), *Rynek dóbr luksusowych w Polsce. Edycja 2014*.
10. Książek M. (2013*), Analiza danych jakościowych*, w: *Zaawansowane metody analiz statystycznych*, Frątczak E. (red.), Szkoła Główna Handlowa, Warszawa.
11. Leigh A. (2009), *Top incomes*, w: *The Oxford handbook of economic inequality*, Salverda W., Nolan B., Smeeding T. (red.), Oxford University Press, Oxford.
12. McFadden D. (1977), *Quantitative methods for analyzing travel behaviour of individuals: Some recent developments*, Cowles Foundation Discussion Paper No. 474, Yale University, New Haven.
13. Peichl A., Schaefer T., Scheicher C. (2008), *Measuring richness and poverty: A micro data application to Europe and Germany*, IZA Discussion Papers No. 3790, Institute for the Study of Labor (IZA).

14. Rada Monitoringu Społecznego (2014), *Diagnoza społeczna 2000–2013: zintegrowana baza danych*, http://www.diagnoza.com, dostęp dnia 9.11.2014.
15. Radziukiewicz M. (2006), *Zasięg ubóstwa w Polsce*, PWE, Warszawa.
16. R Development Core Team (2015), *R: a language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, http://www.r-project.org.
17. Rusnak Z. (2012), *Logistic regression model in poverty analyses*, „Ekonometria" nr 1.
18. Stanisz A. (2007), *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny*, t. 2, *Modele liniowe i nieliniowe*, StatSoft, Kraków.
19. *Top incomes over the twentieth century* (2007), Atkinson A., Piketty T. (red.), Oxford University Press, Oxford.
20. Więckowska B (2015), *Podręcznik użytkownika – PQStat*, PQStat Software.
21. Żarnowski J. (1992), *Bieda i dostatek 1918–1939*, w: *Nędza i dostatek na ziemiach polskich od średniowiecza po wiek XX*, Sztetyłła J. (red.), Seria: Instytut Historii Kultury Materialnej PAN, Semper, Warszawa.

Streszczenie

Celem artykułu była identyfikacja czynników objaśniających bogactwo gospodarstw domowych. W analizie zastosowano model logitowy, w którym rolę zmiennej zależnej pełniła zmienna binarna – przynależność do sfery bogactwa, przyjmująca wartość jeden, gdy gospodarstwo domowe należało do sfery bogactwa oraz wartość zero, gdy gospodarstwo domowe nie należało do sfery bogactwa. Wśród potencjalnych czynników uwzględniono zarówno cechy głowy gospodarstwa domowego (np. płeć, wiek, wykształcenie), jak i cechy samego gospodarstwa (np. miejsce zamieszkania, liczba osób). Oszacowany model poddano weryfikacji statystycznej polegającej na badaniu statystycznej istotności parametrów oraz na określeniu stopnia dopasowania modelu do danych empirycznych.

Słowa kluczowe

bogactwo, determinanty bogactwa dochodowego, model regresji logistycznej

Identification of determinants of income richness using logistic regression model (Summary)

The aim of the paper was identifying the factors explaining income richness. The logit model in which the dependent variable was binary was used – variable equals to 1 if household was rich and equals to 0 if household was not

rich. Among the potential factors there were taken into account characteristics of household (e.g. place of resident, number of persons in household) and household's head (e.g. gender, age, education). The goodness of fit and statistical significance of estimated parameters were evaluated.

Keywords
richness, determinants of income richness, logistic regression model