DOI: 10.5604/01.3001.0014.1160

PRZEGLĄD STATYSTYCZNY
R. LXIII – ZESZYT 2 – 2016

ANNA SĄCZEWSKA-PIOTROWSKA[1]

ZASTOSOWANIE KRZYWYCH ROC W ANALIZIE UBÓSTWA MIEJSKICH I WIEJSKICH GOSPODARSTW DOMOWYCH

1. WPROWADZENIE

W analizie ubóstwa są stosowane często modele logitowe i probitowe, które pozwalają określić czynniki (dotyczące np. cech głowy gospodarstwa domowego) mające statystycznie istotny wpływ na wzrost lub spadek szans pobytu w sferze ubóstwa. Również w literaturze polskiej do określenia czynników ubóstwa znajdują zastosowanie modele logitowe (por. np. Kasprzyk, Fura, 2011; Rusnak, 2012) i probitowe (np. Panek, Czapiński, 2015).

W niniejszym opracowaniu uwaga zostanie skupiona na modelach logitowych. Dokładność ich dopasowania jest przeważnie przeprowadzana z zastosowaniem testu ilorazu wiarygodności czy testu Hosmera-Lemeshowa. Dosyć rzadko w tym celu są wyznaczane krzywe ROC (ang. *relative operating characteristics*) oraz obliczane pola pod krzywymi ROC, oznaczane jako AUC (ang. *area under curve*). Należy podkreślić, że w innych dyscyplinach ROC i AUC mają szersze zastosowanie i są narzędziem często wykorzystywanym do oceny poprawności klasyfikatora, a klasyfikator ten nie jest ograniczony jedynie do modelu logitowego. Przykładowo, w medycynie krzywe ROC są stosowane do porównania nowej metody diagnostycznej (rola klasyfikatora) do obowiązującego „złotego standardu". W przypadku oceny włóknienia wątroby „złotym standardem" jest ocena histopatologiczna preparatu wątroby uzyskanego z oligobiopsji, a metodą porównywaną z tym standardem jest sprężystość miąższu wątroby oceniona na podstawie elastografii. Krzywe ROC i pola AUC wspomagają tym samym system decyzyjny w różnych dziedzinach życia, a najczęściej są stosowane w psychologii, medycynie czy nauczaniu maszynowym. Pionierską pracę dotyczącą zastosowania krzywych ROC i pól AUC w analizie ubóstwa opublikował Wodon (1997). W swojej pracy porównał krzywe ROC i pola AUC wyznaczone dla modeli logitowych uwzględniających różne zestawy wskaźników służących do identyfikacji ubogich (np. położenie, własność ziemska, poziom wykształcenia głowy gospodarstwa domowego). Swoją analizę przeprowadził na przykładzie Bangladeszu. W porównaniach wykorzystał fakt, że położenie jednej krzywej ROC nad drugą krzywą ROC świadczy o tym, że dany

[1] Uniwersytet Ekonomiczny w Katowicach, Wydział Ekonomii, Katedra Metod Statystyczno-Matematycznych w Ekonomii, ul. 1 Maja 50, 40-287 Katowice, Polska, e-mail: anna.saczewska-piotrowska@ue.katowice.pl.

wskaźnik lepiej identyfikuje ubogich. Krzywe ROC w analizie ubóstwa stosowali później m.in. Baulch (2002) oraz Houssou, Zeller (2009).

Niniejsze opracowanie jest próbą przeniesienia dotychczasowych rozważań dotyczących zastosowania krzywych ROC w analizie ubóstwa na grunt polski. Celem opracowania jest zastosowanie krzywych ROC i pól AUC do oceny, który z wyznaczonych modeli logitowych ryzyka ubóstwa – dla gospodarstw w mieście i dla gospodarstw na wsi – ma większą moc predykcyjną. Zmiennymi objaśniającymi w modelach są cechy gospodarstwa domowego (np. liczba osób w gospodarstwie, status gospodarstwa na rynku pracy) oraz jego głowy (np. płeć i wykształcenie). Krzywe ROC pełnią również rolę pomocniczą w wyznaczaniu punktu odcięcia, czyli takiego poziomu prawdopodobieństwa, poniżej którego gospodarstwo uznawane jest za nieubogie (w pakietach statystycznych domyślny punkt odcięcia to 0,5). Dzięki wyznaczeniu krzywych ROC i pól AUC dla uproszczonych modeli logitowych, określono wskaźniki najlepiej identyfikujące ubogie gospodarstwa domowe w mieście i na wsi.

W pracy celowo nie skupiono się na problemach metodologicznych dotyczących pomiaru ubóstwa, odsyłając do literatury przedmiotu. Zjawisko ubóstwa występuje z różną siłą i różnym natężeniem na całym świecie, przyczynia się do powstawania lub pogłębiania innych problemów, np. jest jedną z zasadniczych przyczyn wykluczenia społecznego, i z tego powodu szczegółowy opis problemów metodologicznych pomiaru ubóstwa jest szeroko opisany w literaturze przedmiotu. W Polsce problemy związane z pomiarem ubóstwa omawia szczegółowo Panek (2011). Należy nadmienić, że problem pojawia się już na etapie definiowania tego zjawiska i wiąże się z podjęciem decyzji, czy ubóstwo będzie rozumiane w sposób absolutny czy relatywny, w sposób obiektywny czy subiektywny oraz w sposób klasyczny czy wielowymiarowy. W ujęciu absolutnym gospodarstwa domowe są ubogie, gdy ich podstawowe potrzeby nie są zaspokajane na minimalnym akceptowalnym poziomie. Koncepcja ubóstwa relatywnego zawiera w sobie odniesienie do sytuacji przeciętnej, czyli sytuacji innych gospodarstw domowych. W ujęciu subiektywnym oceny poziomu zaspokojenia potrzeb dokonują same badane gospodarstwa domowe, natomiast w przypadku ujęcia obiektywnego ocena poziomu zaspokojenia potrzeb badanych gospodarstw jest dokonywana niezależnie od ich osobistych wartościowań w tym zakresie. Ubóstwo rozumiane w sposób klasyczny jest postrzegane w kategoriach pieniężnych (przez pryzmat dochodów lub wydatków), natomiast ubóstwo rozumiane w sposób wielowymiarowy jest dodatkowo postrzegane przez pryzmat zasobów materialnych (np. dobra trwałego użytku, mieszkanie itd.) ocenianych w formie niemonetarnej. W przeprowadzonej analizie przyjęto, że ubóstwo jest rozumiane obiektywnie przez pryzmat dochodów z uwzględnieniem odniesienia do sytuacji innych gospodarstw domowych. Po dokonaniu wyboru definicji ubóstwa, kolejne problemy decyzyjne pojawiają się w związku z określeniem wskaźnika zamożności (przyjęto dochody netto gospodarstw domowych), granicy ubóstwa (60% mediany rozkładu dochodów ekwiwalentnych) oraz skal ekwiwalentności (zmodyfikowana skala OECD).

2. DWUMIANOWY MODEL LOGITOWY

Model logitowy może być modelem dwumianowym lub wielomianowym. Dwumianowego modelu logitowego 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):

$$p_i = P(Y_i = 1 \mid x_{i1}, x_{i2}, ..., x_{ik}) = \frac{\exp\left(a_0 + \sum_{j=1}^{k} a_j x_{ik} \right)}{1 + \exp\left(a_0 + \sum_{j=1}^{k} a_j x_{ik} \right)}, \tag{1}$$

gdzie a_j ($j = 0,1,\ldots,k$) są parametrami. Model (1) jest więc modelem wiążącym prawdopodobieństwo jednego z dwóch możliwych wyników zmiennej Y ze zmiennymi objaśniającymi. Współczynniki regresji są zazwyczaj estymowane metodą największej wiarygodności (MNW). Wartości $\exp(a_j)$ w modelu (1) są najczęściej interpretowane przy pomocy pojęcia ilorazu szans (ang. *odds ratio*). Szansa jest definiowana jako relacja prawdopodobieństwa wystąpienia zdarzenia do prawdopodobieństwa niewystąpienia zdarzenia. Wyrażenie $\exp(a_0)$ jest równe szansie dla grupy referencyjnej, tzn. grupy, w której wszystkie zmienne objaśniające są równe zero. W przypadku zmiennej dychotomicznej X_i iloraz szans pokazuje, ilu krotnie zmienia się szansa u jednostki, dla której $X_i = 1$ względem jednostki, dla której $X_i = 0$, przy niezmienionych wartościach pozostałych zmiennych objaśniających. 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).

Estymatory MNW mają asymptotyczny rozkład normalny. Z tego powodu test istotności dla pojedynczego parametru wykorzystuje statystykę z o rozkładzie $N(0,1)$ (Gruszczyński, 2012). Do testowania statystycznej istotności wszystkich parametrów przy zmiennych objaśniających stosuje się test ilorazu wiarygodności (tzw. LR test). W teście LR hipoteza zerowa głosi, że wszystkie parametry przy zmiennych są równe zero, natomiast hipoteza alternatywna, że przynajmniej jeden z nich jest różny od zera. Statystyka ilorazu wiarygodności jest określona wzorem (Gruszczyński, 2001, s. 64; Książek, 2013, s. 60–61):

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

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 uporządkowanych według wartości dopasowanych z modelu logitowego (najczęściej dla grup decylowych) porównuje obserwowane liczebności i oczekiwane liczebności występowania wartości wyróżnionej (zmienna objaśniana przyjmująca wartość 1). 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):

$$HL = \sum_{g=1}^{G} \frac{\left(O_g - E_g\right)^2}{E_g\left(1 - \frac{E_g}{N_g}\right)}, \tag{3}$$

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 testu Hosmera-Lemeshowa brak podstaw do odrzucenia hipotezy zerowej jest pożądany, ponieważ wskazuje na podobieństwo liczebności obserwowanych i oczekiwanych.

Miarą dopasowania modelu jest również miara zaproponowana przez McFaddena (tzw. pseudo-R^2) określona wzorem (McFadden, 1977):

$$R^2_{McFadden} = 1 - \frac{\ln L_{FM}}{\ln L_0}. \tag{4}$$

Pseudo-R^2 bazuje na porównaniu wartości funkcji wiarygodności w oszacowanym modelu i modelu bez zmiennych objaśniających. Miara ta przyjmuje wartości z zakresu $[0,1)^2$, 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 McFadden (1977) wartości z zakresu 0,2–0,4 świadczą o bardzo dobrym dopasowaniu modelu do danych.

3. CZUŁOŚĆ, SPECYFICZNOŚĆ, KRZYWE ROC I POLE AUC

Często najważniejszą miarą dopasowania w modelach logitowych jest ich zdolność predyktywna. Należy podkreślić, że 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.

[2] Miernik ten może przyjąć wartość maksymalną 1 tylko w przypadku, gdy wektor wartości zmiennych objaśniających jest różny dla każdej badanej jednostki (Hosmer, Lemeshow, 2000, s. 166).

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:

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

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):

$$\hat{y}_i = \begin{cases} 1 \text{ dla } \hat{p}_i > p^*, \\ 0 \text{ dla } \hat{p}_i \leq p^*. \end{cases} \tag{6}$$

gdzie p^* jest nową wartością odcinającą (ang. *cut-off point*), wyznaczoną dla danej próby oraz dla danego badania. Dla wybranego punktu odcięcia można zbudować tablicę trafności prognoz (tabela 1).

Tabela 1.

Tablica trafności prognoz

Obserwowane wartości zmiennej objaśnianej	Przewidywane wartości zmiennej objaśnianej		
	$\hat{y}_i = 0$	$\hat{y}_i = 1$	Razem
$y_i = 0$	n_{00}	n_{01}	$n_{0\bullet}$
$y_i = 1$	n_{10}	n_{11}	$n_{1\bullet}$
Razem	$n_{\bullet 0}$	$n_{\bullet 1}$	n

Źródło: opracowanie własne na podstawie Sompolska-Rzechuła i inni (2014).

Na podstawie tabeli 1 można obliczyć następujące mierniki (Gruszczyński, 2001, s. 83–84; Davis, Goadrich, 2006; Dudek, Dybciak, 2006; Jackowska, Wycinka, 2009; Harańczyk, 2010):

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

$$ACC = \frac{n_{00} + n_{11}}{n}, \tag{7}$$

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

$$SE = \frac{n_{11}}{n_{1\bullet}}, \tag{8}$$

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

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

$$SP = \frac{n_{00}}{n_{0\bullet}}, \tag{9}$$

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

4. Wartość predykcyjna dodatniego wyniku (ang. *positive predictive value*, *precision*) określająca udział trafnie przewidzianych przez model „jedynek" w ogólnej grupie przewidzianych „jedynek":

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

5. Wartość predykcyjna ujemnego wyniku (ang. *negative predictive value*) będąca proporcją obserwacji trafnie przewidywanych przez model „zer" w ogólnej liczbie przewidzianych „zer":

$$NPV = \frac{n_{00}}{n_{\bullet 0}}. \tag{11}$$

Czułość i specyficzność są najczęściej wykorzystywanymi miarami i są podstawą w konstrukcji krzywych ROC. Jeżeli mamy do czynienia z wieloma obserwacjami, to wtedy mamy również wiele tablic trafności do przeanalizowania. Aby wybrać tablicę z najlepszym podziałem, warto wykorzystać krzywe ROC 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 AUC i traktować je 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):

$$Z = \frac{\hat{AUC} - 0{,}5}{\sqrt{\hat{Var}(\hat{AUC})}}, \tag{12}$$

gdzie $\hat{Var}(\hat{AUC})$ jest estymatorem wariancji pola $\hat{AUC}$. 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).

W przypadku pól AUC przyjmuje się ogólną zasadę dotyczącą oceny jakości klasyfikacyjnej modeli (Hosmer, Lemeshow, 2000, s. 162; Kumari, Rajnish, 2015):
$AUC = 0{,}5$ klasyfikacja nie jest dobra (jest porównywalna z rzutem monetą),
$0{,}5 < AUC < 0{,}6$ słaba klasyfikacja,
$0{,}6 \leq AUC < 0{,}7$ akceptowalna klasyfikacja,
$0{,}7 \leq AUC < 0{,}8$ dobra klasyfikacja,
$0{,}8 \leq AUC < 0{,}9$ bardzo dobra klasyfikacja,
$AUC \geq 0{,}9$ wyśmienita/wybitna klasyfikacja.

Powyższe rozważania dotyczące czułości i specyficzności oraz krzywych ROC można odnieść do ubóstwa (Wodon, 1997). Czułość *SE* jest frakcją gospodarstw domowych poza sferą ubóstwa, które zostały prawidłowo zaklasyfikowane przez model. Oznaczając przez P, P^- oraz P^+ liczbę gospodarstw ubogich, liczbę gospodarstw ubogich zaklasyfikowanych jako nieubogie oraz liczbę gospodarstw ubogich prawidłowo zaklasyfikowanych jako ubogie, $SE = P^+ / (P^- + P^+) = P^+ / P$. Specyficzność *SP* jest frakcją gospodarstw domowych zaobserwowanych jako nieubogie i prawidłowo zaklasyfikowanych jako nieubogie. Oznaczając gospodarstwo nieubogie *NP*, mamy $SP = NP^- / (NP^- + NP^+) = NP^- / NP$. Nieprawidłowe klasyfikacje oznaczymy przez $1 - SP$ (frakcja gospodarstw zaobserwowanych jako nieubogie i zaklasyfikowanych jako ubogie) i $1 - SE$ (udział gospodarstw zaobserwowanych jako ubogie i zaklasyfikowanych jako nieubogie). W dalszej części będziemy mówić odpowiednio o błędach *SP* i *SE* (tabela 2).

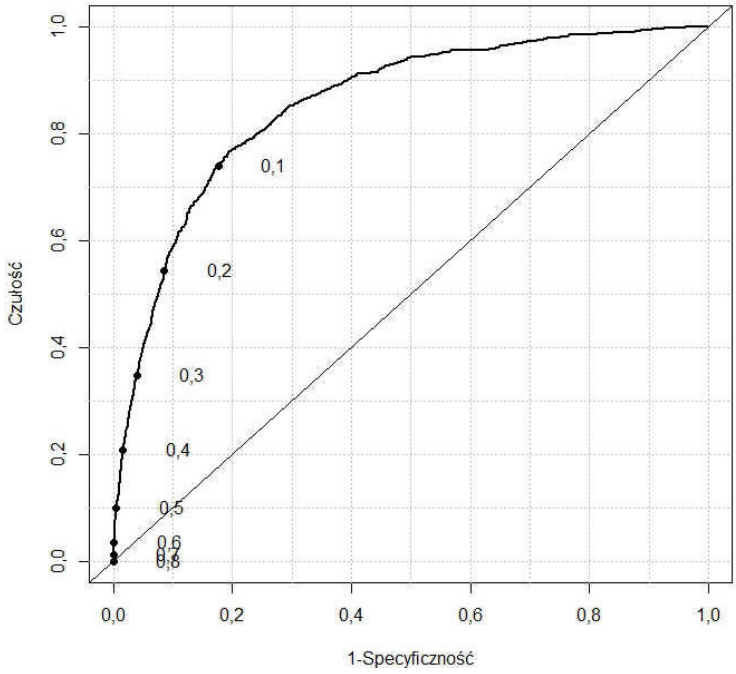
Tabela 2.

Czułość, specyficzność, błędy *SP* i *SE*

Obserwowane wartości	Przewidziane wartości	
	Nieubogie gospodarstwo	Ubogie gospodarstwo
Nieubogie gospodarstwo	$SP = NP^- / NP$	$1 - SP = NP^+ / NP$
Ubogie gospodarstwo	$1 - SE = P^- / P$	$SE = P^+ / P$

Źródło: opracowanie własne na podstawie Wodon (1997).

W przypadku wzrostu wartości punktu odcięcia, mniej gospodarstw domowych będzie uważanych za ubogie, specyficzność *SP* będzie rosła, a wartość błędu *SP* będzie malała. Z drugiej strony czułość *SE* będzie malała, a wartość błędu *SE* będzie rosła. Spadek wartości punktu odcięcia wiąże się z tym, że więcej gospodarstw domowych będzie przewidywanych jako ubogie, *SE* wzrośnie, a wartość błędu *SE* zmaleje, ale wartość błędu *SP* wzrośnie. Krzywe ROC przedstawiają błędy *SP* i *SE* uzyskane w kontinuum punktów odcięcia. Przykładowa krzywa ROC została przedstawiona na rysunku 1.



Rysunek 1. Przykładowa krzywa ROC
Źródło: opracowanie własne.

W początku wykresu (0,0), odpowiadającym punktowi odcięcia $p^* = 1$, *SE* wynosi zero, a *SP* przyjmuje wartość równą jeden. Przy $p^* = 1$ żadne gospodarstwo domowe nie jest klasyfikowane jako ubogie, a prawdopodobieństwo, że gospodarstwo nieubogie będzie zaklasyfikowane jako ubogie wynosi zero. Stąd wartość błędu *SP* musi być równa zero. Ponadto, przy $p^* = 1$ prawdopodobieństwo, że biedne gospodarstwo domowe zostanie zaklasyfikowane jako nieubogie wynosi jeden. Wartość błędu *SE* jest wtedy równa jeden. W odróżnieniu od zaprezentowanej sytuacji, w górnym prawym rogu wykresu (punkt (1,1)), czułość *SE* jest równa jeden, a specyficzność *SP* zero. Odpowiada to punktowi odcięcia $p^* = 0$, wartości błędu *SP* wynoszącemu zero i wartości błędu *SE* wynoszącemu jeden. Pomiędzy tymi dwiema skrajnymi warto-

ściami krzywa ROC przedstawia wartości tych dwóch rodzajów błędów dla różnych wartości p^*.

Im lepiej zastosowany model przewiduje prawdziwe statusy gospodarstw domowych (ubogie lub nieubogie gospodarstwo domowe), tym bardziej krzywa ROC jest bardziej wygięta w stronę lewego górnego rogu wykresu. Jeśli model przewiduje idealnie status przynależności do sfery ubóstwa, jego krzywa ROC przechodzi przez punkt (0,1), a pole pod krzywą AUC jest równe jeden.

Do przewidywania statusów przynależności do sfery ubóstwa można stosować modele z różnymi zmiennymi objaśniającymi. Dla każdego modelu można wyznaczyć krzywą ROC. W przypadku, gdy jeden model ma krzywą ROC leżącą ponad krzywą wyznaczoną dla innego modelu, to oznacza, że pierwszy model dominuje (majoryzuje) drugi model. W takiej sytuacji, niezależnie od wyboru punktu odcięcia, pierwszy model ma mniejsze wartości błędów *SP* i *SE* niż drugi model. Sytuacja wygląda inaczej, gdy krzywe przecinają się. Załóżmy, że dwie krzywe ROC przecinają się i przykładowo krzywa modelu 1 leży powyżej krzywej modelu 2 tylko w dolnej części wykresu. Dolna część wykresu odpowiada wysokim wartościom punktu odcięcia, niskim wartościom błędów *SP* i wysokim wartościom błędów *SE*. Jeżeli decydenci są bardziej wyrozumiali na ryzyko identyfikowania jako nieubogich tych gospodarstw domowych, które są w rzeczywistości ubogie, wybiorą model 1. Jeżeli natomiast dla decydentów mniejsze znaczenie ma ryzyko zaklasyfikowania jako biednego gospodarstwa domowego będącego w rzeczywistości poza sferą ubóstwa, wybiorą wtedy model 2. W przedstawionej sytuacji żaden z modeli nie dominuje innego we wszystkich punktach odcięcia.

4. WIELOWYMIAROWE WSKAŹNIKI

Analizę determinant ubóstwa dochodowego przeprowadzono dla 2015 r. z wykorzystaniem danych projektu „*Diagnoza społeczna*". W badaniu wzięło udział prawie 11 tys. gospodarstw domowych. Jako granicę ubóstwa przyjęto 60% mediany rozkładu dochodów ekwiwalentnych, przy czym zastosowano zmodyfikowaną skalę OECD, zgodnie z którą przypisuje się pierwszej dorosłej osobie w gospodarstwie wartość 1, każdej następnej dorosłej osobie w gospodarstwie wartość 0,5, natomiast wartość 0,3 dziecku (każda osoba poniżej 14 lat). Zmienną zależną w modelu logitowym była zmienna zero-jedynkowa:

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

Zmienne niezależne były zmiennymi jakościowymi, które przedstawiono w postaci układów zmiennych zero-jedynkowych w taki sposób, że zmienna mająca m kategorii jest reprezentowana przez $m - 1$ zmiennych zero-jedynkowych (w ten sposób uniknięto liniowej zależności pomiędzy zmiennymi objaśniającymi). W modelu uwzględnione

zostały zmienne dotyczące płci, wieku i wykształcenia głowy gospodarstwa domowego, liczby osób w gospodarstwie, grupy społeczno-ekonomicznej, statusu gospodarstwa na rynku pracy oraz obecności w gospodarstwie osób z niepełnosprawnością. Mamy tu do czynienia z wielowymiarowymi wskaźnikami (ang. *targeting indicator*, co dosłownie oznacza wskaźnik celowniczy) pomagającymi zidentyfikować gospodarstwa ubogie (wskaźniki te odnoszą się do różnych cech gospodarstwa i jego głowy).

Wszystkie obliczenia i wykresy wykonano w programie R (R Development Core Team, 2015) z wykorzystaniem pakietów `gmodels` (Warnes i inni, 2015), `lmtest` (Hothorn i inni, 2015), `OptimalCutpoints` (Lopez-Raton, Rodriguez-Alvarez, 2015), `ResourceSelection` (Lele i inni, 2015) i `verification` (NCAR – Research Applications Laboratory, 2015). Wymienione pakiety zostały użyte do obliczeń pól AUC i graficznej prezentacji krzywych ROC (`verification`), budowy tablic trafności (`gmodels`), przeprowadzenia testów LR (`lmtest`) i Hosmera-Lemeshowa (`ResourceSelection`) oraz wyboru punktów odcięcia (`OptimalCutpoints`).

W pierwszym kroku oszacowano interesujące z punktu widzenia niniejszego opracowania modele logitowe ryzyka ubóstwa odrębnie dla miejskich (model 1) i wiejskich (model 2) gospodarstw domowych. Wyniki estymacji przedstawiono w tabelach 3 i 4.

Tabela 3.

Wyniki estymacji modelu logitowego ryzyka ubóstwa dla gospodarstw domowych w mieście (model 1)

Zmienne	Współczynnik	Iloraz szans
Stała	-1,992***	x
Płeć głowy gospodarstwa domowego:		
mężczyzna	-0,406***	0,666
kobieta	ref.	
Wiek głowy gospodarstwa domowego:		
34 lata i mniej	ref.	
35–44 lata	0,668**	1,950
45–59 lat	0,483*	1,620
60 i więcej lat	-0,443*	0,642
Wykształcenie głowy gospodarstwa domowego:		
gimnazjum i niższe	ref.	
zawodowe	-0,354**	0,702
średnie	-0,829***	0,436
wyższe	-2,451***	0,086
Liczba osób w gospodarstwie domowym:		
1	ref.	
2	-0,395**	0,674
3	-0,409*	0,664
4	-0,238	0,788
5	-0,001	0,999
6 i więcej	0,017	1,017

Zmienne	Współczynnik	Iloraz szans
Grupa społeczno-ekonomiczna:		
pracownicy	ref.	
rolnicy	0,710	2,034
pracujący na własny rachunek	-1,038**	0,354
emeryci	0,498**	1,645
renciści	1,199***	3,316
utrzymujący się z niezarobkowych źródeł	2,271***	9,693
Status gospodarstwa na rynku pracy:		
przynajmniej jedna osoba bezrobotna	1,720***	5,586
brak osób bezrobotnych	ref.	
Osoby z niepełnosprawnością w gospodarstwie:		
przynajmniej jedna osoba z niepełnosprawnością	0,446***	1,562
brak osób z niepełnosprawnością	ref.	

Ref. – kategoria stanowiąca punkt odniesienia, . $p < 0,1$; * $p < 0,05$; ** $p < 0,01$; *** $p < 0,001$.

Źródło: opracowanie własne na podstawie Rada Monitoringu Społecznego (2015).

Tabela 4.

Wyniki estymacji modelu logitowego ryzyka ubóstwa dla gospodarstw domowych na wsi (model 2)

Zmienne	Współczynnik	Iloraz szans
Stała	-0,847***	x
Płeć głowy gospodarstwa domowego:		
mężczyzna	-0,373***	0,688
kobieta	ref.	
Wiek głowy gospodarstwa domowego:		
34 lata i mniej	ref.	
35–44 lata	0,808***	2,244
45–59 lat	0,258	1,295
60 i więcej lat	-0,130	0,878
Wykształcenie głowy gospodarstwa domowego:		
gimnazjum i niższe	ref.	
zawodowe	-0,443***	0,642
średnie	-0,900***	0,407
wyższe	-2,336***	0,097
Liczba osób w gospodarstwie domowym:		
1	ref.	
2	-0,529***	0,589
3	-0,745***	0,475
4	-0,655***	0,519
5	-0,856***	0,425
6 i więcej	-0,610***	0,543

Tabela 4. (cd.)

Zmienne	Współczynnik	Iloraz szans
Grupa społeczno-ekonomiczna:		
pracownicy	ref.	
rolnicy	0,912***	2,489
pracujący na własny rachunek	-0,590*	0,554
emeryci	0,363**	1,438
renciści	1,368***	3,927
utrzymujący się z niezarobkowych źródeł	1,656***	5,241
Status gospodarstwa na rynku pracy:		
przynajmniej jedna osoba bezrobotna	1,429***	4,173
brak osób bezrobotnych	ref.	
Osoby z niepełnosprawnością w gospodarstwie:		
przynajmniej jedna osoba z niepełnosprawnością	0,189*	1,208
brak osób z niepełnosprawnością	ref.	

Ref. – kategoria stanowiąca punkt odniesienia, . $p < 0,1$; * $p < 0,05$; ** $p < 0,01$; *** $p < 0,001$.

Źródło: opracowanie własne na podstawie Rada Monitoringu Społecznego (2015).

W modelu 1 nieistotne statystycznie zmienne to: liczba osób w gospodarstwie domowym 4 i więcej (w porównaniu do gospodarstw 1-osobowych) oraz przynależność do gospodarstw domowych rolników (w porównaniu do gospodarstw domowych pracowników), natomiast w modelu 2 nieistotne są wiek głowy gospodarstwa 45–59 lat oraz wiek 60 i więcej lat (w porównaniu do gospodarstw z głową 34 lata i mniej).

Wszystkie oszacowane modele poddano weryfikacji, której wyniki zawarto w tabeli 5.

Tabela 5.

Zestawienie wyników weryfikacji oszacowanych modeli logitowych

Wyszczególnienie	Model 1	Model 2
$R^2_{McFadden}$	0,242	0,150
LR test:		
liczba stopni swobody	19	19
χ^2	974,240	762,529
wartość p	0,000	0,000
Test Hosmera-Lemeshowa:		
liczba stopni swobody	8	8
χ^2	14,779	13,864
wartość p	0,064	0,085

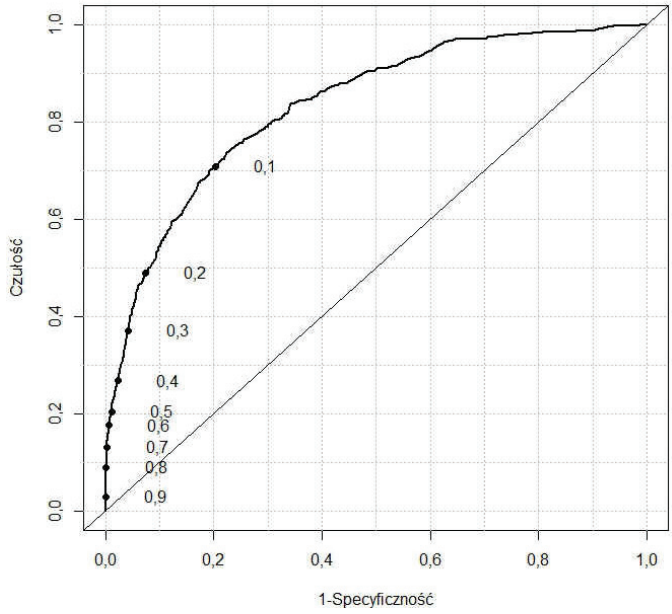
Źródło: opracowanie własne na podstawie Rada Monitoringu Społecznego (2015).

Pseudo-R^2 informuje, że model 1 jest bardzo dobrze dopasowany do danych (miernik przyjmuje wartości z przedziału 0,2–0,4). Na podstawie testu Hosmera-Lemeshowa można stwierdzić, że liczebności obserwowane i teoretyczne nie różnią się istotnie w grupach decylowych w przypadku obydwu oszacowanych modeli. W przypadku obydwu modeli test ilorazu wiarygodności wskazuje, że przynajmniej jeden z parametrów istotnie różni się od zera. Na podstawie przeprowadzonej oceny dopasowania można stwierdzić, że lepiej dopasowany do danych empirycznych jest model 1 (wartość pseudo-R^2 mieści się w odpowiednim przedziale, brak podstaw do odrzucenia hipotezy o różnicy liczebności obserwowanych i teoretycznych oraz odrzucenie hipotezy głoszącej, że wszystkie parametry w tym modelu przyjmują wartość zero).

Analizując ilorazy szans w tabelach 3 i 4 można stwierdzić, że szansa pobytu gospodarstwa domowego w sferze ubóstwa była:

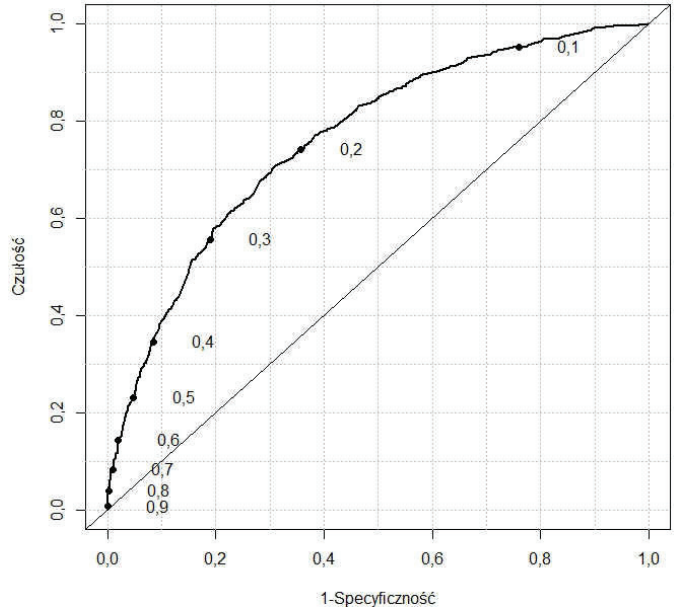
- niższa o ponad 30% w gospodarstwach domowych mężczyzn w porównaniu z gospodarstwami kobiet,
- wyższa w gospodarstwach domowych, których głowa ma 35–44 lata lat niż w gospodarstwach, których głowa ma do 34 lat; należy zaznaczyć, że w miastach szanse ubóstwa są istotnie mniejsze w gospodarstwach, których głowa ma 60 i więcej lat w porównaniu do gospodarstw najmłodszych,
- niższa w gospodarstwach, których głowa ma wykształcenie co najmniej zawodowe w porównaniu do wykształcenia co najwyżej gimnazjalnego, przy czym szanse pobytu w sferze ubóstwa były zdecydowanie najmniejsze w przypadku wykształcenia wyższego (ponad 90% mniejsze szanse bycia ubogim niż w przypadku wykształcenia gimnazjalnego i niższego),
- niższa w gospodarstwach 2- i 3-osobowych w porównaniu do gospodarstw 1-osobowych,
- zdecydowanie niższa w gospodarstwach pracujących na własny rachunek oraz dużo wyższa w gospodarstwach utrzymujących się z niezarobkowych źródeł (w mieście ponad 9-krotnie, na wsi 5-krotnie) niż w gospodarstwach pracowników,
- wyższa kilkukrotnie (w mieście ponad 5-krotnie, na wsi ok. 4-krotnie) w gospodarstwach z przynajmniej jedną osobą bezrobotną niż w gospodarstwach bez osób bezrobotnych,
- wyższa (w mieście o ponad 50%, na wsi o ok. 20%) w gospodarstwach z przynajmniej jedną osobą z niepełnosprawnością w porównaniu do gospodarstw bez osób z niepełnosprawnością.

Dla oszacowanych modeli wyznaczono krzywe ROC (rysunki 2 i 3).



Rysunek 2. Krzywa ROC dla modelu 1

Źródło: opracowanie własne na podstawie Rada Monitoringu Społecznego (2015).



Rysunek 3. Krzywa ROC dla modelu 2

Źródło: opracowanie własne na podstawie Rada Monitoringu Społecznego (2015).

Można zauważyć, że obie krzywe ROC leżą powyżej przekątnej, prostej Czułość = 1-Specyficzność, a tym samym pola AUC są większe niż 0,5. Bardziej wygięta w kierunku punktu (0,1) jest krzywa wyznaczona dla modelu 1. W kolejnym kroku obliczono pola AUC i przeprowadzono test, który pozwolił ocenić, czy pola AUC są istotnie większe niż 0,5 (tabela 6).

Tabela 6.

Pola AUC dla oszacowanych modeli

Wyszczególnienie	Model 1	Model 2
AUC	0,834	0,763
wartość *p*	0,000	0,000

Źródło: opracowanie własne na podstawie Rada Monitoringu Społecznego (2015).

Przeprowadzone testy potwierdzają, że AUC są istotnie większe niż 0,5, co oznacza, że jakość klasyfikacyjna modeli jest dobra i mogą one służyć do budowy prognoz. Jakość klasyfikacyjna modelu 1 jest bardzo dobra ($0{,}8 \leq AUC < 0{,}9$), natomiast modelu 2 jest dobra ($0{,}7 \leq AUC < 0{,}8$). Wartości pól AUC potwierdzają wcześniej wyciągnięte wnioski (na podstawie pseudo-R^2 oraz testów LR i Hosmera-Lemeshowa) o lepszym dopasowaniu modelu 1.

Oszacowane modele cechują się dobrą lub bardzo dobrą jakością klasyfikacyjną, co oznacza, że mogą służyć do przeprowadzania prognoz. Podstawą przeprowadzenia prognoz jest wyznaczenie odpowiednich punktów odcięcia. Badana próba nie była zbilansowana – zdecydowanie więcej gospodarstw (zarówno w mieście jak i na wsi) było poza sferą ubóstwa niż w sferze ubóstwa. Jako punkty odcięcia wybrano punkty krzywych ROC położone[3] najbliżej punktu (0,1), czyli w przypadku modelu 1 była to wartość 0,093, natomiast w przypadku modelu 2 – wartość 0,225. Dla wyznaczonych punktów odcięcia obliczono liczbę poprawnie prognozowanych przypadków (tabele 7 i 8).

[3] Jako punkty odcięcia wyznaczono dodatkowo częstości występowania ubogich gospodarstw domowych – w przypadku gospodarstw miejskich punkt 0,1, natomiast w przypadku gospodarstw wiejskich – punkt 0,235. Uzyskano następujące rezultaty: specyficzność 79,73%, czułość 70,85%, zliczeniowy R^2 78,83% w przypadku gospodarstw miejskich, natomiast specyficzność 70,20%, czułość 69,20%, zliczeniowy R^2 69,96% w przypadku gospodarstw wiejskich.

Tabela 7.

Klasyfikacja przypadków na podstawie modelu 1 dla punktu odcięcia 0,093

Obserwowane wartości zmiennej objaśnianej	Przewidywane wartości zmiennej objaśnianej		
	$\hat{y}_i = 0$	$\hat{y}_i = 1$	Razem
$y_i = 0$	4283	1276	5559
$y_i = 1$	160	461	621
Razem	4443	1737	6180
Specyficzność 77,05%, czułość 74,24%, zliczeniowy R^2 76,76%			

Źródło: opracowanie własne na podstawie Rada Monitoringu Społecznego (2015).

Tabela 8.

Klasyfikacja przypadków na podstawie modelu 2 dla punktu odcięcia 0,225

Obserwowane wartości zmiennej objaśnianej	Przewidywane wartości zmiennej objaśnianej		
	$\hat{y}_i = 0$	$\hat{y}_i = 1$	Razem
$y_i = 0$	2475	1085	3560
$y_i = 1$	327	767	1094
Razem	2802	1852	4654
specyficzność 69,52%, czułość 70,11%, zliczeniowy R^2 69,66%			

Źródło: opracowanie własne na podstawie Rada Monitoringu Społecznego (2015).

Wykorzystując dane z tabel 7 i 8, obliczono specyficzność, czułość oraz zliczeniowy R^2, które pozwoliły ocenić procentową trafność prognoz. Dla modelu 1 procent prawidłowych predykcji wyniósł 76,76%, przy czym specyficzność wyniosła 77,05%, a czułość 74,24%. Można na tej podstawie sądzić, że model ten przewiduje w nieco lepszym stopniu pobyt poza sferą ubóstwa (77,05% nieubogich gospodarstw domowych zostało uznane przez model jako nieubogie) niż pobyt w ubóstwie (74,24% ubogich gospodarstw zostało przewidziane przez model jako ubogie). Odmienna sytuacja jest w przypadku modelu 2, gdzie czułość jest wyższa niż specyficzność, przy czym procent prawidłowych predykcji ogółem oraz procent prawidłowych predykcji pobytu w ubóstwie i poza sferą ubóstwa jest niższy niż dla modelu 1.

Na podstawie modelu 1 zbudowano przykładowe prognozy stanu przynależności do sfery ubóstwa gospodarstw domowych o różnych cechach:

– głowa gospodarstwa to mężczyzna mający 40 lat z wykształceniem wyższym, 2-osobowe gospodarstwo pracowników bez osób bezrobotnych i bez osób z niepełnosprawnością:
  prognozowane prawdopodobieństwo wynosi $\hat{p}_1 = 0{,}01$, czyli na podstawie przyjętego punktu odcięcia $p^* = 0{,}093$ można się spodziewać, że gospodarstwo będzie poza sferą ubóstwa $\hat{y}_1 = 0$,

– głowa gospodarstwa domowego to kobieta w wieku 30 lat z wykształceniem średnim, 3-osobowe gospodarstwo rolników z jedną osobą bezrobotną i z jedną osobą z niepełnosprawnością:
prawdopodobieństwo wynosi $\hat{p}_2 = 0{,}412$, czyli gospodarstwo będzie należeć do sfery ubóstwa ($\hat{y}_2 = 1$).

5. JEDNOWYMIAROWE WSKAŹNIKI

Modele logitowe ryzyka ubóstwa dla gospodarstw miejskich oraz dla gospodarstw wiejskich (modele 1 i 2) można uprościć, ograniczając się do uwzględnienia podzbiorów determinant[4] ubóstwa odnoszących się do wieku głowy gospodarstwa, grupy społeczno-ekonomicznej itd. Wyznaczone podzbiory determinant są jednowymiarowymi wskaźnikami identyfikującymi ubogie gospodarstwa domowe (dotyczą każdorazowo jednej cechy gospodarstwa domowego lub jego głowy). Dla każdego uproszczonego modelu opartego na podzbiorze determinant można wyznaczyć krzywą ROC oraz obliczyć pole pod wykresem krzywej AUC.

W tabeli 9 przedstawiono wyniki obliczeń pól pod wykresami krzywych ROC dla uproszczonych modeli.

Tabela 9.

Pola pod wykresami krzywymi ROC dla podzbiorów determinant ubóstwa

Podzbiór determinant	Gospodarstwa domowe			
	miasto		wieś	
	AUC	wpływ	AUC	wpływ
Wykształcenie głowy gospodarstwa	0,688	1	0,628	2
Grupa społeczno-ekonomiczna	0,668	2	0,653	1
Status gospodarstwa na rynku pracy	0,643	3	0,592	3
Liczba osób w gospodarstwie	0,574	5	0,567	4
Osoby z niepełnosprawnością	0,580	4	0,537	6
Płeć głowy gospodarstwa	0,564	7	0,563	5
Wiek głowy gospodarstwa	0,572	6	0,524	7
Wszystkie determinanty	0,834	-	0,763	-

Źródło: opracowanie własne na podstawie Rada Monitoringu Społecznego (2015).

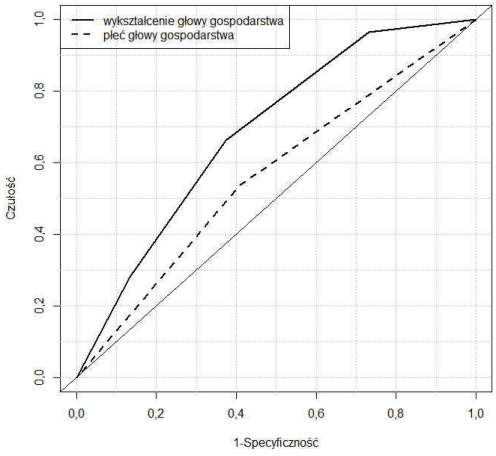
[4] Pojedynczy podzbiór determinant tworzy w naszym przypadku zmienna jakościowa z wszystkimi kategoriami. Może się zdarzyć, że więcej niż jedna zmienna będzie się odnosić do jednego wymiaru i wtedy podzbiór determinant będzie obejmował wszystkie zmienne wraz z ich kategoriami.

Wszystkie pola AUC są statystycznie istotne większe od 0,5 (wartość $p = 0{,}000$), niemniej jednak uwzględnienie w modelu wskaźników odnoszących się do jednego wymiaru nie daje zadowalających rezultatów klasyfikacyjnych (AUC nie przekracza 0,7, co oznacza co najwyżej akceptowalną klasyfikację).

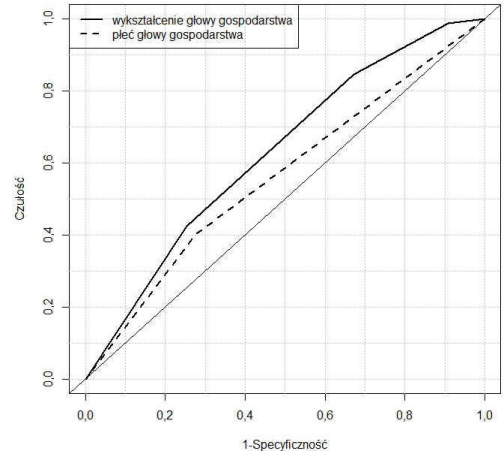
Wyznaczone pola AUC wskazują, że największe znaczenie w identyfikacji gospodarstw ubogich ma wykształcenie głowy gospodarstwa domowego (gospodarstwa miejskie) i grupa społeczno-ekonomiczna (gospodarstwa wiejskie). Najgorszymi wskaźnikami są natomiast płeć (gospodarstwa w mieście) oraz wiek głowy gospodarstwa domowego (gospodarstwa na wsi).

Wyznaczenie krzywych ROC dla każdego uproszczonego modelu pozwoli odpowiedzieć na pytanie, czy w grupach gospodarstw miejskich i wiejskich istnieją podzbiory wskaźników dominujące inne podzbiory. Kierując się jedynie wynikami obliczeń AUC można stwierdzić, że wykształcenie głowy gospodarstwa (gospodarstwa w miastach) i grupa społeczno-ekonomiczna (gospodarstwa wiejskie) są najlepszymi wskaźnikami. Należy jednak zaznaczyć, że są to jedynie miary ogólne. Aby sprawdzić, czy wskaźniki te są najlepsze dla wszystkich możliwych punktów odcięcia należy wyznaczyć krzywe ROC i porównać je ze sobą. Warto zaznaczyć, że w przypadku płci głowy gospodarstwa domowego, obecności osób bezrobotnych oraz niepełnosprawnych w gospodarstwie, zmienne te przyjmowały po dwie kategorie, co oznacza, że do wyznaczenia krzywej ROC (poza punktami (0,0) i (1,1)) wykorzystuje się tylko jeden punkt. W przypadku np. wykształcenia głowy gospodarstwa występują cztery kategorie, stąd krzywą ROC wyznacza się za pomocą trzech punktów (poza punktami (0,0) i (1,1)).

a) gospodarstwa w mieście



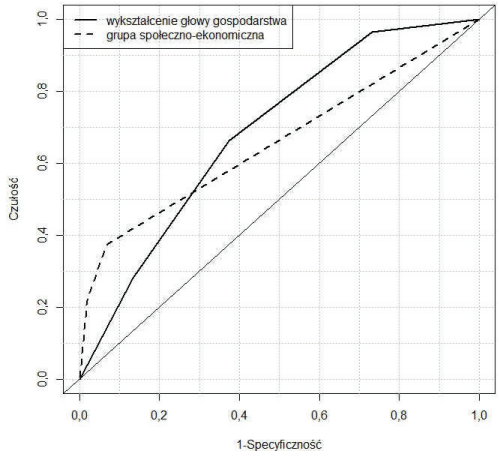
b) gospodarstwa na wsi



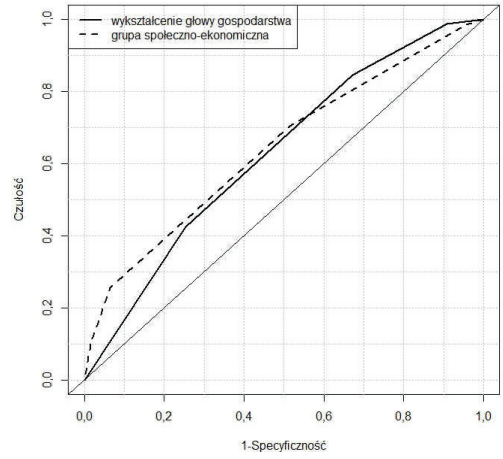
Rysunek 4. Krzywe ROC dla wykształcenia i płci głowy gospodarstwa domowego
Źródło: opracowanie własne na podstawie Rada Monitoringu Społecznego (2015).

Na podstawie rysunku 4 można stwierdzić, że w grupie gospodarstw miejskich i wiejskich wykształcenie głowy dominuje płeć głowy gospodarstwa domowego. Wniosek ten można sformułować na podstawie tego, że wyznaczone krzywe ROC nie przecinają się, a krzywe dla podzbioru determinant odnoszących się do wykształcenia głowy leżą nad krzywymi wyznaczonymi dla płci głowy gospodarstwa. Przykładowo, nie można wskazać dominujących wskaźników w parze wykształcenie głowy gospodarstwa domowego i grupa społeczno-ekonomiczna, ponieważ wyznaczone krzywe ROC przecinają się (rysunek 5). W gospodarstwach domowych w mieście i na wsi lepszym wskaźnikiem dla niskich wartości punktów odcięcia jest wykształcenie głowy gospodarstwa domowego, natomiast grupa społeczno-ekonomiczna jest lepszym wskaźnikiem dla wysokich wartości punktów odcięcia.

a) gospodarstwa w mieście



b) gospodarstwa na wsi



Rysunek 5. Krzywe ROC dla wykształcenia głowy gospodarstwa domowego i grupy społeczno-ekonomicznej

Źródło: opracowanie własne na podstawie Rada Monitoringu Społecznego (2015).

Przeprowadzając porównania krzywych ROC dla wszystkich uproszczonych modeli w grupach gospodarstw w mieście i na wsi można zauważyć, że w przypadku gospodarstw miejskich grupa społeczno-ekonomiczna dominuje kilka wskaźników: odnoszących się do niepełnosprawności członków gospodarstwa, liczby osób w gospodarstwie domowym oraz płci i wieku głowy gospodarstwa domowego. Wyznaczona dla modelu uwzględniającego grupę społeczno-ekonomiczną ogólna miara AUC pozwoliła na wyciągnięcie odmiennego wniosku (najlepszy wskaźnik to wykształcenie głowy gospodarstwa), ponieważ miara ta nie bierze pod uwagę wielkości błędów *SP* i *SE* dla wszystkich punktów odcięcia. W grupie gospodarstw wiejskich dwoma dominującymi wskaźnikami są grupa społeczno-ekonomiczna i status gospodarstwa na rynku

pracy. Dominują one cztery wskaźniki odnoszące się do: niepełnosprawności, liczby osób oraz wieku i płci głowy gospodarstwa. Do wskazanej przez miarę AUC grupy społeczno-ekonomicznej dołączył więc inny wskaźnik – status gospodarstwa na rynku pracy. Model uwzględniający ten wskaźnik ma bowiem, niezależnie od wyboru punktu odcięcia, mniejsze wartości błędów *SE* i *SP* w porównaniu do modeli uwzględniających wymienione cztery wskaźniki.

6. PODSUMOWANIE

Krzywe ROC i pola AUC są stosunkowo rzadko stosowane w analizie zjawisk ekonomicznych. Wykorzystanie krzywych ROC oraz pól AUC ogranicza się zazwyczaj do oceny poprawności klasyfikatora oraz wyboru optymalnego punktu odcięcia. W niniejszym opracowaniu przedstawiono również możliwości wykorzystania krzywych ROC do oceny mocy predykcyjnej wskaźników pomagających namierzyć ubogie gospodarstwa domowe. Graficznie, zbiór wskaźników dominuje inny zbiór wskaźników jeśli krzywa ROC wyznaczona dla jednego modelu leży nad krzywą wyznaczoną dla drugiego modelu. Dodatkowo, pola AUC obydwu modeli dostarczają przydatnej statystyki podsumowującej ich moce predykcyjne.

Na podstawie przeprowadzonej analizy można stwierdzić, że oszacowane modele logitowe ryzyka ubóstwa uwzględniające zmienne odnoszące się do cech gospodarstwa domowego i jego głowy cechują się wysoką dobrocią dopasowania, na co wskazują zarówno „tradycyjne" miary (pseudo-R^2, test wiarygodności ilorazu, test Hosmera-Lemeshowa) oraz wyznaczone krzywe ROC i obliczone pola AUC. Dla każdego z uproszczonych modeli logitowych odnoszących się podzbiorów determinant (jednowymiarowe wskaźniki) również wyznaczono krzywe ROC oraz obliczono pola AUC, co pozwoliło stwierdzić, że najlepsze rezultaty w identyfikacji ubogich gospodarstw domowych w mieście i na wsi są osiągane przy zastosowaniu wskaźnika odnoszącego się do grupy społeczno-ekonomicznej. W obydwu grupach gospodarstw domowych wskaźnik ten dominował kilka innych wskaźników odnoszących się do obecności osób niepełnosprawnych w gospodarstwie, liczby osób w gospodarstwie oraz do płci i wieku głowy gospodarstwa domowego.

Podjęcie decyzji dotyczącej zaklasyfikowania gospodarstwa domowego do jednej z dwóch grup – gospodarstw ubogich lub nieubogich – jest analogiczne do decyzji podejmowanej przez bank przy udzielaniu kredytu – klient dobry i zły. W pierwszym przypadku zaklasyfikowanie gospodarstwa do grupy ubogich łączy się z wypłatą świadczeń lub udzieleniem pomocy rzeczowej, natomiast w drugim przypadku – zaklasyfikowanie klienta do grupy dobrych klientów wiąże się z wypłatą kredytu. W literaturze dotyczącej ryzyka kredytowego (np. Osiewalski, 2007; Marzec, 2008) pojawiają się nowe propozycje wyznaczania prawdopodobieństwa granicznego p^* oraz są stosowane coraz bardziej zaawansowane modele zmiennych jakościowych, co może stanowić wskazówkę kierunku dalszych badań nad trafnością podejmowania decyzji dotyczących wsparcia gospodarstw domowych żyjących w ubóstwie.

LITERATURA

Baulch B., (2002), *Poverty Monitoring and Targeting Using ROC Curves: Examples from Vietnam*, IDS Working Paper 161, Institute of Development Studies, Brighton.

Davis J., Goadrich M., (2006), *The Relationship between Precision-recall and ROC Curves*, Technical Report No. 1551, Computer Sciences Department, University of Wisconsin, Madison.

Dudek H., Dybciak M., (2006), Zastosowanie modelu logitowego do analizy wyników egzaminu, *Zeszyty Naukowe SGGW, Ekonomika i Organizacja Gospodarki Żywnościowej*, 60.

Gruszczyński M., (2001), *Modele i prognozy zmiennych jakościowych w finansach i bankowości*, SGH, Warszawa.

Gruszczyński M., (2012), Modele zmiennych jakościowych dwumianowych, w: Gruszczyński M., (red.), *Mikroekonometria. Modele i metody analizy danych indywidualnych*, Wolters Kluwer, 71–122.

Harańczyk G., (2010), Krzywe ROC, czyli ocena jakości klasyfikatora i poszukiwanie optymalnego punktu odcięcia, w: *Medycyna i analiza danych*, StatSoft, Kraków, 79–89.

Hosmer D. W., Lemeshow S., (2000), *Applied Logistic Regression*, John Wiley & Sons, Hoboken.

Hothorn T., Zeileis A., Farebrother R. W., Cummins C., Millo G., Mitchell D., (2015), *Lmtest: Testing Linear Regression Models*, https://cran.r-project.org/web/packages/lmtest/index.html.

Houssou N., Zeller M., (2009), *Targeting the Poor and Smallholder Farmers: Empirical Evidence from Malawi*, Department of Agricultural Economics and Social Sciences in the Tropics and Subtropics, Discussion Paper No. 1/2009.

Jackowska B., (2011), Efekty interakcji między zmiennymi objaśniającymi w modelu logitowym w analizie zróżnicowania ryzyka zgonu, *Przegląd Statystyczny*, 58 (1–2), 24–41.

Jackowska B., Wycinka E., (2009), Modele ryzyka skreślenia z listy studentów na przykładzie studentów trybu niestacjonarnego, w: Jajuga K., Walesiak M., (red.), *Taksonomia 16. Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, 7, UE, Wrocław, 485–493.

Kasprzyk B., Fura B., (2011), Wykorzystanie modeli logitowych do identyfikacji gospodarstw domowych zagrożonych ubóstwem, *Wiadomości Statystyczne*, 6 (601), 1–16.

Kopczewska K., Kopczewski T., Wójcik P., (2009), *Metody ilościowe w R. Aplikacje ekonomiczne i finansowe*, CeDeWu, Warszawa.

Książek M., (2013), Analiza danych jakościowych, w: Frątczak E., (red.), *Zaawansowane metody analiz statystycznych*, SGH, Warszawa, s. 25–138.

Kumari D., Rajnish K., (2015), Comparing Efficiency of Software Fault Prediction Models Developed through Binary and Multinomial Logistic Regression Techniques, w: Mandal J. K., Satapathy S. C., Sanyal M. K., Sarkar P. P., Mukhopadhyay A., (red.), *Information Systems Design and Intelligent Applications*, Proceedings of Second International Conference India 2015, 1, Springer, 87–198.

Lele S. R., Keim J. L., Solymos P., (2015), *ResourceSelection: Resource Selection (Probability) Functions for Use-availability Data*, https://cran.r-project.org/web/packages/ResourceSelection/index.html.

Lopez-Raton M., Rodriguez-Alvarez M. X., (2015), *OptimalCutpoints: Computing Optimal Cutpoints in Diagnostic Tests*, https://cran.r-project.org/web/packages/OptimalCutpoints/index.html.

Marzec J. (2008), *Bayesowskie modele zmiennych jakościowych i ograniczonych w badaniach niespłacalności kredytów*, UE, Kraków.

McFadden D., (1977), *Quantitative Methods for Analyzing Travel Behaviour of Individuals: Some Recent Developments*, Cowles Foundation Discussion Paper No. 474, Yale University, New Haven.

NCAR – Research Applications Laboratory, (2015), *Verification: Weather Forecast Verification Utilities*, https://cran.r-project.org/web/packages/verification/index.html.

Osiewalski J., (2007), Bayesowska statystyka i teoria decyzji w analizie ryzyka kredytu detalicznego, w: Fatuła D., (red.), *Finansowe warunkowania decyzji ekonomicznych*, Krakowska Szkoła Wyższa, Kraków, 15–28.

Panek T., (2011), *Ubóstwo, wykluczenie społeczne i nierówności. Teoria i praktyka pomiaru*, SGH, Warszawa.

Panek T., Czapiński J., (2015), Wykluczenie społeczne. Diagnoza społeczna 2015. Warunki i jakość życia Polaków – raport, *Contemporary Economics*, 9 (4), 396–432.
R Development Core Team, (2015), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, http://www.r-project.org.
Rada Monitoringu Społecznego, (2015), *Diagnoza społeczna 2000–2015: zintegrowana baza danych*, http://www.diagnoza.com (28.12.2015 r.).
Rusnak Z., (2012), Logistic Regression Model in Poverty Analyses, *Ekonometria*, 1 (35), 9–23.
Sompolska-Rzechuła A., Machowska-Szewczyk M., Chudecka-Głaz A., Cymbaluk-Płoska A., Menkiszak J., (2014), The Use of Logistic Regression in the Ovarian Cancer Diagnostics, *Ekonometria*, 3 (45), 151–164.
Stanisz A., (2007), *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny*, Tom 2, Modele liniowe i nieliniowe, StatSoft, Kraków.
Warnes G. R., Bolker B., Lumley T., Johnson R. C., (2015), *Gmodels: Various R Programming Tools for Model Fitting*, https://cran.r-project.org/web/packages/gmodels/index.html.
Więckowska B., (2015), *Podręcznik użytkownika – PQStat*, PQStat Software.
Wodon Q. T., (1997), Targeting the Poor Using ROC Curves, *World Development*, 25 (12), 2083–2092.

ZASTOSOWANIE KRZYWYCH ROC W ANALIZIE UBÓSTWA MIEJSKICH I WIEJSKICH GOSPODARSTW DOMOWYCH

Streszczenie

W artykule przeprowadzono analizę determinant ubóstwa miejskich i wiejskich gospodarstw domowych z wykorzystaniem dwumianowego modelu logitowego. Do oceny mocy predykcyjnej oszacowanych modeli ryzyka ubóstwa gospodarstw miejskich i wiejskich zastosowano krzywe ROC oraz pola pod krzywymi ROC, oznaczane jako AUC. Wyboru punktów odcięcia (poziom prawdopodobieństwa, poniżej którego gospodarstwo uznawane jest za nieubogie) dokonano na podstawie wyznaczonych krzywych ROC. Dzięki wyznaczeniu krzywych ROC i pól AUC dla uproszczonych modeli logitowych, czyli zawierających podzbiory determinant określono wskaźnik najlepiej identyfikujący ubogie gospodarstwa domowe w mieście i na wsi. Najlepsze rezultaty w identyfikacji ubogich gospodarstw w mieście i na wsi są osiągane przy zastosowaniu wskaźnika odnoszącego się do grupy społeczno-ekonomicznej gospodarstwa domowego.

Słowa kluczowe: ubóstwo, model logitowy, identyfikacja ubogich, krzywe ROC, miejskie i wiejskie gospodarstwa domowe

APPLICATION OF ROC CURVES IN POVERTY ANALYSIS OF URBAN AND RURAL HOUSEHOLDS

Abstract

The article analyses poverty determinants of urban and rural households using binomial logit model. There were used ROC curves and area under ROC curves (AUC) to evaluate the predictive power of estimated risk models of urban and rural households poverty. Based on ROC curves there were chosen cut-off points (level of probability below which a household is considered not poor). On the basis of ROC curves and areas under ROC curves for simplified logit models (containing subsets of determinants) there was pointed the best poverty indicator. In urban and in rural areas the best targeting indicator is socio-economic group of household.

Keywords: poverty, logit model, targeting the poor, ROC curves, urban and rural households