PRZEGLĄD STATYSTYCZNY
R. LXII– ZESZYT 1 – 2015

ANNA SĄCZEWSKA-PIOTROWSKA

BADANIE UBÓSTWA Z ZASTOSOWANIEM NIEPARAMETRYCZNEJ ESTYMACJI FUNKCJI PRZEŻYCIA DLA ZDARZEŃ POWTARZAJĄCYCH SIĘ

1. WSTĘP

Ubóstwo jest przeważnie rozpatrywane w ujęciu statycznym przez pryzmat dochodów lub wydatków gospodarstwa domowego. Oznacza to, że na podstawie danych dotyczących dochodów lub wydatków w danym momencie czasowym, obliczane są wybrane mierniki ubóstwa, dotyczące najczęściej jego zasięgu, głębokości czy dotkliwości. Można zauważyć, że w Polsce i za granicą coraz częściej są przeprowadzane badania panelowe, które pozwalają poszerzyć badania nad ubóstwem o wymiar czasowy. Dynamika i trwałość ubóstwa mogą być analizowane z wykorzystaniem metod analizy historii zdarzeń, zwanych również metodami analizy przeżycia lub metodami analizy trwania. Analiza dynamiki ubóstwa z zastosowaniem tych metod została zapoczątkowana przez Bane, Ellwood (1986), przy czym skupiali się oni w badaniach na czasach oczekiwania na pierwsze wyjście ze sfery ubóstwa oraz na pierwsze wejście do sfery ubóstwa. W trakcie okresu obserwacji gospodarstwa domowe mogą jednak wchodzić i wychodzić ze sfery ubóstwa wielokrotnie. W późniejszych badaniach Stevens (1994, 1999) podkreślała, że w badaniach należy uwzględniać wszystkie epizody (epizod to czas oczekiwania na wystąpienie zdarzenia, czyli na wejście lub wyjście ze sfery ubóstwa), ponieważ tuż po wyjściu ze sfery ubóstwa gospodarstwa domowe są bardziej narażone na ponowne wejście do sfery ubóstwa. Z tego powodu, zdaniem Stevens, rozkład ogółu lat spędzonych w sferze ubóstwa w przypadku analizy wielokrotnych epizodów jest zupełnie inny niż w przypadku analizy pojedynczych epizodów. Należy również zwrócić uwagę na sugestie innych autorów (np. Allison, 2010), którzy podkreślają, że w przypadku, gdy średnia liczba epizodów przypadająca na jednostkę jest niewielka (mniejsza niż dwa), to analizę lepiej ograniczyć jedynie do pierwszego epizodu.

Celem niniejszego opracowania jest analiza czasu trwania gospodarstw domowych w sferze ubóstwa oraz poza sferą ubóstwa. Analiza zostanie przeprowadzona za pomocą nieparametrycznych estymatorów funkcji przeżycia dla powtarzających się epizodów (estymatory Wanga-Changa oraz Peñy-Strawdermana-Hollandera). Należy podkreślić, że estymatory te będą po raz pierwszy wykorzystywane w badaniu trwałości ubóstwa. Wyniki uzyskane za pomocą wymienionych estymatorów zostaną porównane z wynikami uzyskanymi za pomocą znanego estymatora funkcji przeżycia dla pojedynczego epizodu – estymatora Kaplana-Meiera.

2. WYMIAR CZASOWY W BADANIU UBÓSTWA

Problemy z pomiarem ubóstwa pojawiają się już etapie definiowania tego zjawiska, następnie wiążą się z określeniem wskaźnika dobrobytu ekonomicznego, granicy ubóstwa oraz skal ekwiwalentności. Zdefiniowanie kategorii ubóstwa jest pierwszym i najważniejszym krokiem na drodze pomiaru jego charakterystyk. Wszystkie definicje ubóstwa można dopasować do jednej z następujących kategorii:

- ubóstwo to posiadanie mniej niż obiektywnie zdefiniowane absolutne minimum,
- ubóstwo to posiadanie mniej niż inni w społeczeństwie,
- ubóstwo to uczucie, że nie ma się wystarczająco dużo, aby sobie poradzić.

Zgodnie z pierwszą kategorią definicji ubóstwo jest absolutne (bezwzględne), zgodnie z drugą kategorią – relatywne (względne), natomiast według trzeciej kategorii może być absolutne, relatywne lub mieszane. Inna różnica pomiędzy kategoriami polega na tym, że trzecia kategoria definiuje ubóstwo jako subiektywną sytuację, podczas gdy pierwsza i druga kategoria – jako obiektywną sytuację (Hagenaars, de Vos, 1988).

Pomiar ubóstwa absolutnego polega na określeniu wartości dochodów potrzebnych do zakupu dóbr i usług zaspokajających niezbędne potrzeby jednostki. Koncepcja ubóstwa relatywnego zawiera natomiast w sobie odniesienie poziomu zaspokojenia potrzeb jednostek do poziomu ich zaspokojenia przez innych członków społeczeństwa (Panek, 2011, s. 18–19).

W ujęciu subiektywnym oceny poziomu zaspokojenia potrzeb dokonują same badane jednostki (osoby, rodziny, gospodarstwa domowe), natomiast w przypadku ujęcia obiektywnego ocena poziomu zaspokojenia potrzeb badanych jednostek jest dokonywana niezależnie od ich osobistych wartościowań w tym zakresie (Panek i inni, 1999, s. 11).

Badając ubóstwo należy również podjąć decyzję, czy ubóstwo będzie rozumiane w sposób klasyczny czy wielowymiarowy, tzn. czy będzie postrzegane jedynie w kategoriach pieniężnych (przez pryzmat dochodów lub wydatków) czy również przez pryzmat zasobów materialnych (np. dobra trwałego użytku, mieszkanie itd.) ocenianych w formie niemonetarnej.

Szczególnie istotnym podziałem ubóstwa, z punktu widzenia niniejszego opracowania, jest podział na ubóstwo chwilowe i chroniczne (trwałe, długookresowe). Ubóstwo trwałe bywa interpretowane jako „szczególny przypadek ubóstwa wielowymiarowego, w którym czas jest dodatkowym – poza dochodem (konsumpcją) – wymiarem uwzględnianym w analizie" (Topińska, 2008). Trwały pobyt w sferze ubóstwa jest szczególnie niebezpieczny dla jednostek, ponieważ sprzyja wykluczeniu społecznemu oraz degradacji biologicznej. Z tego powodu monitorowanie ubóstwa w dłuższej perspektywie czasu jest niezbędnym elementem skutecznej polityki społecznej państwa.

W badaniach ubóstwa uwzględniających czas są stosowane cztery rodzaje metod (Rodgers, Rodgers, 1993; Layte, Fouarge, 2004), wśród których znajduje się metoda oparta na analizie liczby okresów (ang. *spell-based approach*). Wykorzystuje ona modele trwania do oszacowania czasu trwania w sferze ubóstwa, ryzyka wyjścia ze sfery ubóstwa w zależności od czasu trwania okresu ubóstwa i różnych cech jednostek oraz zajmuje się szacowaniem średniego czasu trwania okresu ubóstwa. Prekursorami tego podejścia byli Bane, Ellwood (1986), którzy brali pod uwagę pojedyncze okresy spędzone w ubóstwie oraz poza sferą ubóstwa. Stevens (1994, 1999) analizując ubóstwo brała pod uwagę wszystkie okresy spędzone przez gospodarstwa domowe w sferze oraz poza sferą ubóstwa. Badaniem ubóstwa z wykorzystaniem metod analizy historii zdarzeń zajmowali się również między innymi Fouarge, Layte (2005), Callens, Croux (2009) oraz Andriopoulou, Tsakloglou (2011). Niewątpliwą zaletą metody opartej na liczbie okresów jest to, że uwzględnia problem cenzurowania obserwacji, tzn. czy początek i koniec sekwencji okresów spędzonych w ubóstwie oraz poza sferą ubóstwa jest znany czy też nie.

Należy zaznaczyć, że wybór określonej definicji ubóstwa implikuje określony sposób pomiaru ubóstwa. Przyjmując definicję ubóstwa np. w ujęciu relatywnym, dysponujemy pewnym wachlarzem możliwości wyboru granicy ubóstwa czy skal ekwiwalentności. Wyczerpujący opis metodologii ubóstwa można znaleźć w literaturze przedmiotu, zarówno zagranicznej, np. Hagenaars, van Praag (1985), Atkinson i inni (2002), jak i krajowej (na pierwszy plan wysuwają się prace T. Panka[1]).

3. ANALIZA POJEDYNCZEGO ZDARZENIA

3.1. ANALIZA HISTORII ZDARZEŃ – PODSTAWOWE POJĘCIA I MIARY

Analiza historii zdarzeń jest ogólnym pojęciem odnoszącym się do grupy statystycznych metod pozwalających analizować czas oczekiwania na wystąpienie zdarzenia (Mills, 2011, s. 257). Przez zdarzenie (ang. *event*) należy rozumieć każdą zmianę w wartościach cechy pierwotnej powodującą przejście z jednego stanu w drugi. Cechy pierwotne identyfikują stan pobytu, natomiast cechy wtórne różnicują między sobą jednostki będące w tym samym stanie pobytu (Frątczak i inni, 2005, s. 22). Czas oczekiwania na wystąpienie zdarzenia T nazywany jest czasem przeżycia, czasem trwania lub epizodem (ang. *survival time*, *duration*, *spell*). Czas przeżycia jest nieujemną zmienną losową wyrażoną w latach, miesiącach, tygodniach, dniach itd. Konkretna wartość T jest oznaczana jako t.

Podstawowe miary w analizie historii zdarzeń są rozpatrywane dla pojedynczego epizodu i kiedy wyróżnia się jeden stan wyjścia (ang. *origin state*) i jeden stan przeznaczenia (ang. *destination state*), mamy do czynienia z populacją homogeniczną,

[1] Przykładowe prace: Panek i inni (1999), Panek (2011).

a zmienna *T* jest zmienną ciągłą. Miary te w literaturze przedmiotu są określane jako miary w analizie pojedynczych epizodów (Frątczak i inni, 2005, s. 37).

Kluczowymi miarami występującymi w analizie trwania są funkcja przeżycia *S*(*t*) oraz funkcja hazardu (ryzyka, intensywności zdarzeń) $\lambda(t)$. Funkcja przeżycia może być wyrażona w odniesieniu do dystrybuanty. Dystrybuanta zmiennej losowej *T* określa prawdopodobieństwo, że czas przeżycia *T* jest mniejszy bądź równy wartości *t* (Hosmer i inni, 2008, s. 16; Mills, 2011, s. 9):

$$\mathrm{F(t)} = \mathrm{P(T} \leq \mathrm{t)}. \tag{1}$$

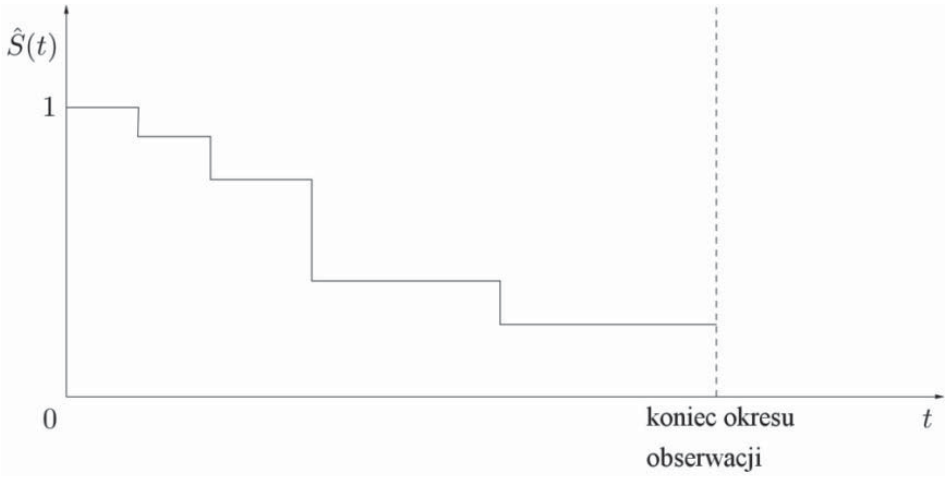
Funkcję przeżycia definiuje się w następujący sposób

$$S(t) = 1 - F(t) = P(T > t). \tag{2}$$

Wyraża ona prawdopodobieństwo tego, że czas przeżycia *T* jest większy od wartości *t* Funkcja przeżycia ma następujące teoretyczne własności (Kleinbaum, Klein, 2005, s. 9): jest nierosnąca, gładka oraz *S*(0) = 1 i $\lim_{t\to\infty} S(t) = 0$. W praktyce, zdarzenia są obserwowane na dyskretnej skali czasu (dni, tygodnie itd.) i z tego powodu wykres funkcji przeżycia jest funkcją schodkową. Wykres często nie zmierza do zera na końcu badania, co jest spowodowane tym, że nie każda jednostka doświadcza zdarzenia. Funkcje przeżycia w teorii i praktyce przedstawiono odpowiednio na rysunkach 1 oraz 2.

Rysunek 1. Teoretyczna funkcja przeżycia

Źródło: opracowanie własne na podstawie (Kleinbaum, Klein, 2005, s. 9).



Rysunek 2. Funkcja przeżycia w praktyce
Źródło: jak rysunek 1.

Funkcja ryzyka to funkcja wyrażona wzorem

$$\lambda(t) = \frac{f(t)}{S(t)}, \tag{3}$$

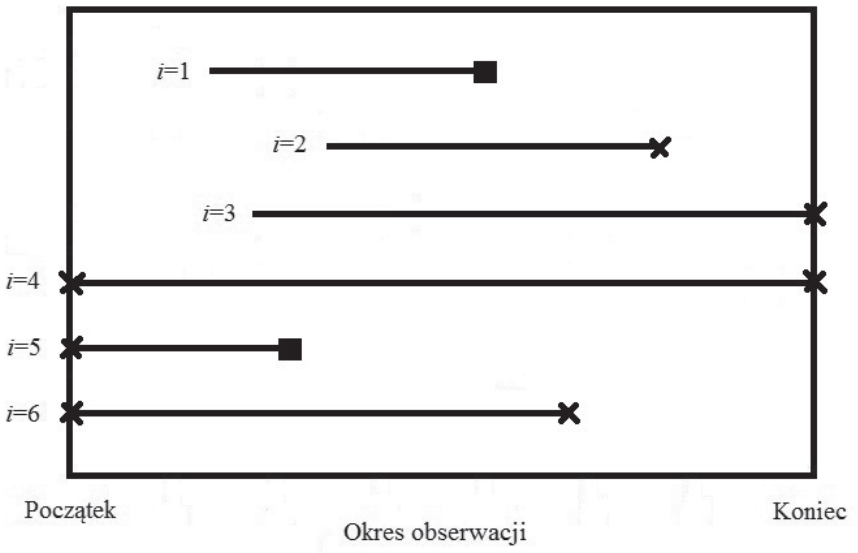
gdzie: $S(t)$ – funkcja przeżycia, f(t) – funkcja gęstości zmiennej czasowej T.

Funkcja ryzyka jest warunkową funkcją gęstości wystąpienia zdarzenia w chwili t, pod warunkiem, że do tej chwili jednostka takiego zdarzenia nie doznała (Bieszk-Stolorz, Markowicz, 2012, s. 23).

3.2. CENZUROWANIE INFORMACJI

W wielu przypadkach historia epizodu nie jest kompletna, co oznacza, że pojawia się problem cenzurowania. Cenzurowanie jest definiowane jako wystąpienie zdarzenia, gdy jednostka nie jest pod obserwacją (Frątczak i inni, 2005, s. 22). Lewostronne cenzurowanie występuje w sytuacji, gdy jednostka doświadcza zdarzenia przed rozpoczęciem okresu obserwacji, natomiast prawostronne cenzurowanie, gdy zdarzenia ma miejsce po zakończeniu okresu obserwacji. Cenzurowanie prawostronne ma miejsce, gdy badanie jest prowadzone przez określony z góry czas, gdy jednostka uczestnicząca w badaniu wycofa się wcześniej lub gdy jednostka znika z pola objętego obserwacją z niewiadomych powodów.

Na rysunku 3 zilustrowano graficznie kilka możliwych sytuacji cenzurowania informacji, przy czym początek linii bez dodatkowego oznaczenia to początek narażenia jednostki na ryzyko wystąpienia zdarzenia, czarny kwadrat symbolizuje wystąpienie zdarzenia, natomiast „x” oznacza cenzurowanie.



Rysunek 3. Cenzurowanie informacji w przypadku pojedynczego epizodu
Źródło: opracowanie własne.

Spośród zaprezentowanych na rysunku sytuacji, tylko w pierwszym przypadku historia epizodu jest kompletna. Epizody drugi i trzeci są cenzurowane prawostronnie. W przypadku epizodu numer dwa można przypuszczać, że jednostka bądź sama zrezygnowała z udziału w dalszych badaniach lub znikła z pola obserwacji z innych powodów. Z taką sytuacją można się często spotkać w przypadku badań panelowych. Cenzurowanie prawostronne epizodu trzeciego rozpoczyna się od końca okresu obserwacji, najczęściej oznacza to datę realizacji badania. Tego typu cenzurowanie może wystąpić w badaniach retrospektywnych i panelowych. Czwarty epizod jest cenzurowany obustronnie, a sytuacja taka może również wystąpić zarówno w badaniach retrospektywnych, jak i panelowych. W przypadku piątego epizodu występuje częściowe obcięcie lewostronne, co oznacza, że nie jest znany początek epizodu oraz pewien jego fragment z początkowej fazy. Ostatni epizod to przykład epizodu cenzurowanego obustronnie, lecz w przeciwieństwie do epizodu czwartego, ucięcie prawostronne ma miejsce wewnątrz okresu obserwacji, a nie na jego końcu.

Należy podkreślić, że rysunek nie wyczerpuje wszystkich rodzajów cenzurowania (np. epizody całkowicie ocenzurowane prawostronnie i lewostronnie), a więcej informacji na temat cenzurowania informacji można znaleźć m.in. w pracy Frątczak i inni (2005).

3.3. KLASY MODELI ANALIZY HISTORII ZDARZEŃ

W analizie przeżycia są stosowane modele nieparametryczne, semiparametryczne oraz parametryczne (opis modeli na podstawie Collett, 2003; Mills, 2011, s. 11–14).

Analizę trwania rozpoczyna się często od modeli nieparametrycznych i właśnie te modele będą stosowane w analizie. Do nieparametrycznych modeli zaliczany jest estymator aktuarialny (metoda tablic trwania życia) oraz estymator Kaplana-Meiera (produktu iloczynowego). W modelach nieparametrycznych nie ma żadnego założenia dotyczącego kształtu funkcji ryzyka oraz wpływu oddziaływania zmiennych objaśniających na kształt tej funkcji, a wpływ zmiennych ujawnia się poprzez rozwarstwienie danych na grupy. Nieparametryczne metody są dobrą metodą do zrozumienia podstaw i dokonania opisu, przy czym estymator aktuarialny jest dobry dla dużych zbiorów danych oraz w przypadku, gdy czasy zdarzeń nie są precyzyjnie mierzone, natomiast estymator Kaplana-Meiera jest bardziej korzystny dla mniejszych prób oraz dla danych mierzonych precyzyjnie.

Do modeli semiparametrycznych należy zaliczyć modele Coxa oraz modele wykładnicze stałe przedziałami. W modelach tych nie ma założenia dotyczącego kształtu funkcji ryzyka, lecz jest silne założenie dotyczące wpływu oddziaływania zmiennych na kształt funkcji hazardu pomiędzy grupami na przestrzeni czasu. Do estymacji parametrów modeli semiparametrycznych stosowana jest metoda częściowej wiarygodności. Nie można stosować tradycyjnej metody największej wiarygodności, ponieważ nie ma wyspecyfikowanej parametrycznie funkcji dla hazardu bazowego (jest to hazard przy założeniu, że wszystkie zmienne objaśniające są równe zero). Modele semiparametryczne są modelami elastycznymi, pozwalającymi uwzględnić wiele zmiennych, a wyniki są często zbliżone do wyników uzyskanych na podstawie modeli parametrycznych, lecz bez restrykcyjnych założeń. Modele te są jednak mało odpowiednie do testowania hipotez o zależności w czasie (tzn. jak hazard zmienia się w czasie) oraz są mniej precyzyjne od modeli parametrycznych.

Do modeli parametrycznych można zaliczyć między innymi modele: wykładniczy, Weibulla, logistyczny, gamma, normalny, komplementarny log-log, log-normalny, Gompertza. W przypadku tych modeli badacz musi określić z góry kształt funkcji ryzyka oraz jak zmienne wpływają na funkcję ryzyka. Parametry modeli parametrycznych są estymowane metodą największej wiarygodności. W porównaniu do modeli semiparametrycznych, oszacowania parametrów tych modeli są bardziej precyzyjne. W modelach parametrycznych można uwzględnić wiele zmiennych objaśniających, zarówno dyskretnych, jak i ciągłych. Modele te określają kształt funkcji hazardu, dzięki czemu można na ich podstawie przeprowadzać predykcję. Istotną wadą modeli parametrycznych jest to, że przy błędnie określonej funkcji ryzyka, oszacowane parametry modelu mogą być poważnie obciążone.

3.4. ESTYMATOR KAPLANA-MEIERA

Najbardziej popularnym estymatorem funkcji przeżycia dla danych niecenzurowanych oraz cenzurowanych prawostronnie jest estymator Kaplana-Meiera (1958) zwany również estymatorem produktu iloczynowego. Dla próby n epizodów, niech

$t_1 < t_2 < \cdots < t_m$ będą uporządkowanymi czasami trwania epizodów oraz $t_0 = 0$. Estymator Kaplana-Meiera przyjmuje postać:

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right), \tag{4}$$

gdzie: d_i – liczba zdarzeń występujących w momencie t_i, n_i – liczba jednostek narażonych na ryzyko wystąpienia zdarzenia bezpośrednio przed czasem t_i (łącznie z cenzurowanymi czasami przeżycia w momencie t_i).

Wariancję estymatora Kaplana-Meiera określa się, stosując najczęściej formułę Greenwooda (1926):

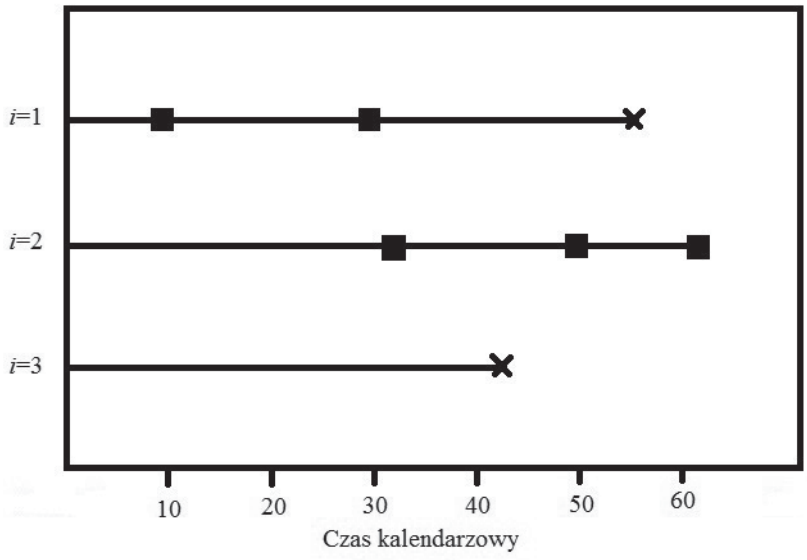
$$\widehat{\text{var}}\left(\hat{S}(t)\right) = \left(\hat{S}(t)\right)^2 \sum_{i:t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}. \tag{5}$$

Peterson (1977) udowodnił, że estymator Kaplana-Meiera jest zgodny, natomiast Breslow i Crowley (1974) pokazali, że $\sqrt{n}\left(\hat{S}(t) - S(t)\right)$ ma asymptotyczny rozkład normalny z wartością oczekiwaną zero oraz macierzą wariancji-kowariancji, która może być aproksymowana z zastosowaniem formuły Greenwooda.

4. ANALIZA ZDARZEŃ POWTARZAJĄCYCH SIĘ

Zdarzenia powtarzające się są zdarzeniami tego samego typu, mogącymi się powtarzać wielokrotnie w ciągu życia jednostki, np. wejście do sfery ubóstwa czy utrata pracy. Na rysunku 4 przedstawiono trzy przykładowe jednostki ze zdarzeniami powtarzającymi się.

Czarny kwadrat symbolizuje wystąpienie zdarzenia, natomiast „x” oznacza cenzurowanie prawostronne. Przyjęto, że informacje nie były cenzurowane lewostronnie, ponieważ każda jednostka spełniała warunek początkowy, jakim było wystąpienie zdarzenia inicjującego (ang. *initial event*). Przykładowo, w analizie czasu trwania ubóstwa zdarzeniem tym jest wejście do sfery ubóstwa. Można zauważyć, że pierwsza jednostka przeżyła dwa zdarzenia przed cenzurowaniem, druga jednostka doświadczyła trzech zdarzeń, kończąc okres obserwacji zdarzeniem (nie występuje cenzurowanie), natomiast jednostka trzecia nie przeżyła żadnego zdarzenia przed cenzurowaniem.



Rysunek 4. Przykład trzech jednostek ze zdarzeniami powtarzającymi się
Źródło: opracowanie własne na podstawie (Kelly, Lim, 2000; González, 2006, s. 5).

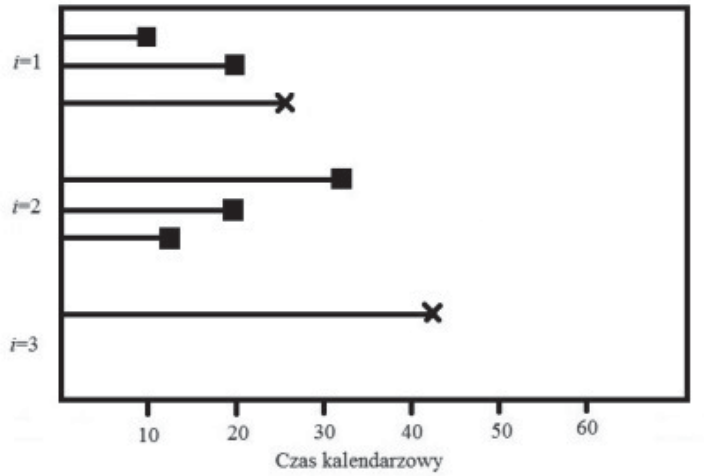
4.1. PRZEDZIAŁY RYZYKA

W przypadku zdarzeń powtarzających się czas jest indeksowany w dwóch skalach: kalendarzowej oraz pomiędzy kolejnymi zdarzeniami. Przedziały ryzyka określają, kiedy jednostka jest narażona na wystąpienie zdarzenia wzdłuż danej skali czasu. Można wyróżnić trzy rodzaje przedziałów ryzyka: luka czasowa, całkowity czas oraz proces liczenia. Wykorzystując dane z rysunku 4, na rysunku 5 zilustrowano rodzaje przedziałów ryzyka.

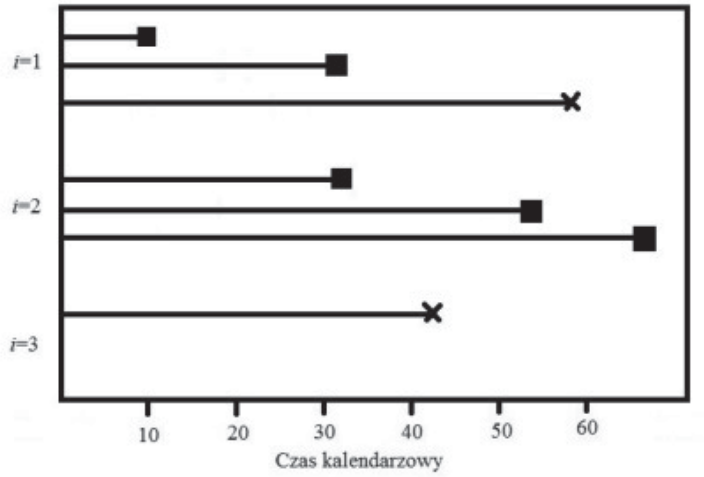
Luka czasowa (ang. *interoccurrence time*, *gap time*) jest czasem odstępu od poprzedniego zdarzenia, co oznacza, że zegar jest restartowany po zajściu każdego zdarzenia. W przypadku całkowitego czasu (ang. *total time*) zegar nie jest restartowany po zajściu zdarzenia, tym samym czas jest liczony od wybranego punktu na skali czasu. Proces liczenia (ang. *counting proces*) wymaga rozważenia czasu kalendarzowego oraz luk czasowych, co oznacza, że jest on procesem podwójnie indeksowanym[2]. Proces liczenia używa tej samej skali czasu jak całkowity czas, ale uwzględnia fakt, że jednostka może mieć opóźnione wejście lub cenzurowany okres przed tym, gdy zaczęła być narażona na ryzyko wystąpienia zdarzenia. W przypadku luki czasowej oraz procesu liczenia jednostka jest zagrożona wystąpieniem zdarzenia przez ten sam okres czasu. Należy zauważyć, że przedział ryzyka dla pierwszego zdarzenia jest taki sam w przypadku wszystkich definicji przedziałów ryzyka.

[2] Proces liczenia jest procesem stochastycznym z dodatnimi, całkowitymi i rosnącymi wartościami.

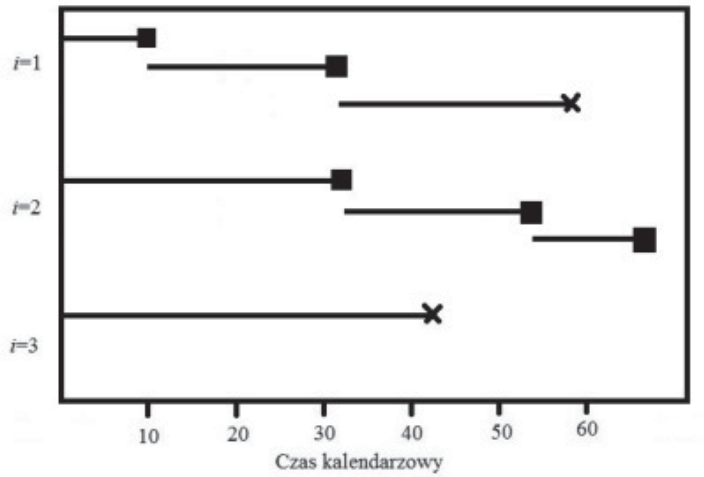
a) luka czasowa



b) całkowity czas



c) proces liczenia



Rysunek 5. Rodzaje przedziałów ryzyka
Źródło: jak rysunek 4.

4.2. DEFINICJE I ZAŁOŻENIA

Luki czasowe mogą być niezależne i mieć takie same rozkłady (ang. *independent and identically distributed* – i.i.d.) lub mogą być ze sobą skorelowane. W pierwszym przypadku mówimy o modelu i.i.d., natomiast w drugim przypadku najczęściej rozpatrywany jest multiplikatywny model słabości (ang. *multiplicative frailty model*) (González, Peña, 2003, 2004; González, 2006).

W modelu i.i.d. zakładamy, że dysponujemy n niezależnymi jednostkami. Dla i-tej jednostki kolejne luki czasowe oznaczamy przez $T_{ij}, j = 1,2,...$, a $T_{i0} = 0$. Zakładamy, że czasy odstępu od poprzedniego zdarzenia są i.i.d. nieujemnymi zmiennymi losowymi z łączną absolutnie ciągłą dystrybuantą $F(t) = P\{T_{ij} \leq t\}$. Ponadto zakładamy, że obserwacja i-tej jednostki ustaje w losowym czasie τ_i, gdzie $\tau_1, \tau_2, ..., \tau_n$ są i.i.d. z łączną dystrybuantą $G(w) = P\{\tau_i \leq w\}$ Zakładamy ponadto, że τ_i oraz T_{ij} są wzajemnie niezależne. Empiryczną dystrybuantę τ_i oznaczmy przez $G_n(t)$. Dla każdego $i = 1,2,...,n$ niech $S_{i0} = 0$ i $S_{ij} = \sum_j T_{ij}, j = 1,2,....$ Liczba wystąpień zdarzenia dla i-tej jednostki jest równa

$$K_i = \max\{j \in \{0,1, ...\}: S_{ij} \leq \tau_i\}. \tag{6}$$

Ostatecznie, dla i-tej jednostki obserwowane są zmienne losowe:

$$\left(K_i, \tau_i, T_{i1}, T_{i2}, ..., T_{iK_i}, \tau_i - S_{iK_i}\right).$$

W przypadku multiplikatywnego modelu słabości zakładamy, że dla każdej jednostki istnieje nieobserwowalna słabość Z_i. Mówiąc, że pewne jednostki są „słabsze", mamy na myśli to, że są bardziej narażone na przeżycie zdarzenia niż inne jednostki. Nieobserwowalna słabość przyjmuje wartości dodatnie. Pod warunkiem $Z_i = z$ luki czasowe $T_{i1}, T_{i2},...$ są i.i.d. z warunkową funkcją przeżycia

$$\bar{F}(t|Z_i = z) = [\bar{F}_0(t)\,]^z = \exp\left(-z\int_0^t \lambda_0(u)du\right), \tag{7}$$

gdzie $\lambda_0(\cdot)$ jest funkcją ryzyka powiązaną z bazową funkcją przeżycia $\bar{F}_0(\cdot)$. Bazowa (podstawowa) funkcja przeżycia jest funkcją, która zależy jedynie od czasu (nie występują w niej inne zmienne objaśniające). Zakłada się, że słabości $Z_1, Z_2, ..., Z_n$ są i.i.d. od nieznanej dystrybuanty H. Słabości nie są obserwowalne, więc szacowane jest brzegowe przeżycie luk czasowych T_{ij}

$$\bar{F}(t) = \mathbf{E}\left\{\bar{F}_0^{Z_1}(t)\right\} = \psi[\Lambda_0(t)], \tag{8}$$

gdzie $\psi(u) = \mathbf{E}\{\exp(-uZ_1)\}$ jest transformatą Laplace'a Z_1, a $\Lambda_0(t) = -\log[F_0(t)]$ jest skumulowaną funkcją ryzyka $\bar{F}_0$. Brzegowa skumulowana funkcja ryzyka jest określona wzorem

$$\Lambda(\mathrm{t}) = -\int_0^{\mathrm{t}} \frac{\psi'\big(\Lambda_0(u)\big)}{\psi\big(\Lambda_0(u)\big)}\lambda_0(u)du = \int_0^{t} \frac{\mathbf{E}\big[\mathrm{Z}_1\bar{F}_0^{Z_1}(u)\big]}{\mathbf{E}\big[\bar{F}_0^{Z_1}(u)\big]}\lambda_0(u)du. \tag{9}$$

Obydwie postaci wynikają bezpośrednio z definicji $\psi(\cdot)$.

Popularnym wyborem nieznanej dystrybuanty słabości H jest dystrybuanta gamma z parametrem kształtu i parametrem skali równym nieznanemu parametrowi α. W tym przypadku łączna brzegowa funkcja przeżycia $\bar{F}$ w (8) przyjmuje postać:

$$\bar{F}(t) = \left[\frac{\alpha}{\alpha + \Lambda_0(t)}\right]^{\alpha}. \tag{10}$$

Parametr α określa stopień skorelowania pomiędzy lukami czasowymi w obrębie jednostki. W szczególności, jeżeli α rośnie (maleje), zależność między czasami odstępu maleje (rośnie). Przyjmując $\alpha \to \infty$ otrzymujemy model z niezależnymi lukami czasowymi, w którym luki T_{ij} mają łączną funkcję przeżycia $\bar{F}_0$. W rzeczywistości, gdy $\alpha \to \infty$ to Z_i zbiega do jedności według prawdopodobieństwa dla $i = 1,2,...,n$.

4.3. ESTYMATOR WANGA-CHANGA

Wang i Chang (1999) zaproponowali estymator łącznej brzegowej funkcji przeżycia zakładający skorelowanie luk czasowych w obrębie jednostek. Autorzy rozważali strukturę korelacji w dość ogólny sposób, uwzględniając jako specjalne przypadki zarówno model i.i.d., jak i model słabości gamma. Przyjmując wszystkie wagi[3] równe jeden (a_i w notacji Wanga i Changa), estymator przyjmuje postać

$$\hat{S}(t) = \prod_{\{T_j \in \tau; T_j \le t\}} \left[1 - \frac{d^*(T_j)}{R^*(T_j)}\right]. \tag{11}$$

Przy konstrukcji estymatora Wanga-Changa ostatnia luka czasowa nie jest brana pod uwagę, chyba że i-ta jednostka nie doświadczyła zdarzenia w czasie $[0, \tau_i]$:

[3] Wagi te zależą od długości czasów τ_i. Na podstawie przeprowadzonych symulacji, Wang i Chang stwierdzili, że przyjęcie jednakowych wag jest bardzo dobrym wyborem.

$$K_i^* = I\{K_i = 0\} + K_i I\{K_i > 0\}. \tag{12}$$

Ponadto, d^* jest sumą proporcji jednostek, których czasy odstępu między zdarzeniami są równe t w momencie wystąpienia zdarzenia:

$$d^*(t) = \sum_{i=1}^{n} \frac{1}{K_i^*} \sum_{j=1}^{K_i} I\{T_{ij} = t\}, \tag{13}$$

natomiast R^* jest średnią liczba jednostek narażonych na ryzyko w czasie t, a τ jest zbiorem różnych zaobserwowanych kompletnych luk czasowych dla n jednostek

$$R^*(t) = \sum_{i=1}^{n} \frac{1}{K_i^*} \left[\sum_{j=1}^{K_i} I\{T_{ij} \geq t\} + I\{\tau_i - S_{iK_i} \geq t\} I\{K_i = 0\} \right]. \tag{14}$$

Asymptotyczna wariancja estymatora Wanga-Changa jest zdefiniowana wzorem:

$$\mathrm{var}\left(\hat{S}(t)\right) = \bar{F}(t)^2 \sigma_{WC}^2(t), \tag{15}$$

gdzie $\sigma_{WC}^2(t)$ zdefiniowali Wang i Chang (w artykule przyjęli oznaczenie ϕ). Asymptotyczna wariancja może być szacowana w następujący sposób:

$$\widehat{\mathrm{var}}\left(\hat{S}(t)\right) = \hat{S}(t)^2 \hat{\sigma}_{WC}^2(t), \tag{16}$$

gdzie:

$$\hat{\sigma}_{WC}^2(t) = \sum_{i=1}^{n} \frac{1}{K_i^*} \sum_{\{T_j \in \tau; T_j \leq t\}} \frac{d^*(T_j)}{R^{*2}(T_j)}. \tag{17}$$

4.4. ESTYMATORY PEÑY-STRAWDERMANA-HOLLANDERA

Peña i inni (2001) odkryli nieparametryczny estymator największej wiarygodności będący uogólnieniem estymatora Kaplana-Meiera przy założeniu, że zdarzenia powtarzają się, a czasy pomiędzy zdarzeniami są niezależne i mają te same rozkłady. Z tego powodu estymator ten nazywany jest IIDPLE (ang. *independent and identically product limit estimator*).

Dla danego czasu kalendarzowego *s* oraz luki czasowej *t*, czyli wykorzystując proces podwójnego indeksowania, definiujemy:

$$K_i(s) = \sum_{j=1}^{\infty} I\{S_{ij} \leq s\}, \tag{18}$$

$$N(s,t) = \sum_{i=1}^{n} \sum_{j=1}^{K_i(s)} I\{T_{ij} = t\}, \tag{19}$$

$$Y(s,t) = \sum_{i=1}^{n} \left[\sum_{j=1}^{K_i(s-)} I\{T_{ij} \geq t\} + I\{\min(s, \tau_i) - S_{iK_i(s-)} \geq t\} \right]. \tag{20}$$

Estymator łącznej funkcji przeżycia $\bar{F}$ czasu między zdarzeniami zaproponowany przez Peñę i inni (2001) jest definiowany następująco:

$$\hat{\bar{F}}(s,t) = \prod_{w \leq t} \left[1 - \frac{N(s, \Delta w)}{Y(s,w)} \right], \tag{21}$$

gdzie: $N(s,t)$ – zlicza ilość zaobserwowanych zdarzeń występujących w okresie kalendarzowym [0, *s*], których czasy odstępu od poprzedniego zdarzenia są równe co najwyżej *t*, $Y(s,t)$ – zlicza ilość zaobserwowanych zdarzeń występujących w okresie kalendarzowym [0, *s*], których czasy odstępu są równe co najmniej *t*.

Gdy luki czasowe są skorelowane w obrębie jednostek, obciążenie IIDPLE jest większe niż w przypadku estymatora Wanga-Changa. W pracy Peñy i inni (2001) można znaleźć szerszą informację dotyczącą porównania właściwości tych estymatorów oraz dowody zgodności i słabej zbieżności estymatora IIDPLE.

Estymator $\hat{\bar{F}}(s,t)$ ma asymptotyczną wariancję, gdy $n \to \infty$ daną wzorem:

$$\operatorname{var}\left(\hat{\bar{F}}(s,t)\right) = \bar{F}(s,t)^2 \sigma^2_{PSH}(s,t). \tag{22}$$

Estymator wariancji jest określony wzorem:

$$\widehat{\operatorname{var}}\left(\hat{\bar{F}}(s,t)\right) = \hat{\bar{F}}(s,t)^2 \hat{\sigma}^2_{PSH}(s,t), \tag{23}$$

gdzie:

$$\hat{\sigma}^2_{PSH}(s,t) = \int_0^t \frac{N(s,dw)}{Y(s,w)[Y(s,w) - N(s,\Delta w)]}. \tag{24}$$

Peña i inni (2001) zaproponowali również estymator łącznej brzegowej funkcji przeżycia luk czasowych w przypadku, gdy są one skorelowane w obrębie jednostek, a korelacja jest indukowana przez model słabości gamma. Estymator ten nazywany jest FRMLE (ang. *frailty maximum likelihood estimator*). Oszacowanie α oraz Λ_0 (wzór 2) może być uzyskane maksymalizując brzegową funkcję wiarygodności z zastosowaniem algorytmu maksymalizacji wartości oczekiwanej (EM). Metoda ta została szczegółowo opisana w pracy Peñy i inni (2001). FRMLE przyjmuje postać:

$$\tilde{\tilde{F}}(s,t) = \left[\frac{\hat{\alpha}}{\hat{\alpha} + \hat{\Lambda}_0(s,t)}\right]^{\hat{\alpha}}, \tag{25}$$

gdzie: $\hat{\alpha}$ – estymator parametru skali α, $\hat{\Lambda}_0(s,t)$ – estymator brzegowej skumulowanej funkcji hazardu.

Można traktować ten estymator jako bezpośrednie uogólnienie estymatora $\tilde{\tilde{F}}(s,t)$ do modelu słabości gamma. Model i.i.d. jest osiągany przyjmując $\alpha \to \infty$ co wymusza Z_i do zbieżności według prawdopodobieństwa do jedności. Z tego powodu można się spodziewać, że $\hat{\tilde{F}}$ oraz $\tilde{\tilde{F}}$ będą bliskie w modelu i.i.d. FRMLE nie posiada przybliżonej postaci dla asymptotycznej wariancji, dlatego niezbędna jest dalsza praca w celu zrozumienia asymptotycznego zachowania $\tilde{\tilde{F}}(s,t)$.

5. ANALIZA CZASU TRWANIA W SFERZE UBÓSTWA I W SFERZE POZA UBÓSTWEM

W badaniu wykorzystano dane pochodzące z siedmiu etapów panelu zrealizowanego w latach 2000–2013 w ramach projektu „Diagnoza społeczna" (Rada Monitoringu Społecznego, 2013). Jako wskaźnik zamożności przyjęto dochody netto gospodarstw domowych w Polsce w lutym/marcu 2000, 2003, 2005, 2007, 2009, 2011 i 2013 r. W celu uwzględnienia różnic występujących w wielkości i składzie demograficznym gospodarstwa domowych obliczono dochody ekwiwalentne, stosując w tym celu zmodyfikowaną skalę OECD. Skala ta przypisuje 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). Dochody ekwiwalentne są ważone liczbą gospodarstw domowych. Gospodarstwo domowe zostało uznane za ubogie, gdy jego dochód był mniejszy niż 60% mediany rozkładów dochodów ekwiwalentnych w danym roku. Należy zaznaczyć, że dokonane wybory dotyczące pomiaru ubóstwa są subiektywne, a zastosowanie innej definicji ubóstwa,

wskaźnika dobrobytu ekonomicznego, granicy ubóstwa oraz skal ekwiwalentności może doprowadzić do uzyskania odmiennych rezultatów dotyczących ubóstwa.

Należy podkreślić, że warunkiem uwzględnienia danego gospodarstwa domowego w analizie czasu trwania w ubóstwie oraz poza ubóstwem było wystąpienie zdarzenia początkowego. Przykładowo, aby rozpocząć analizę czasu trwania ubóstwa, gospodarstwo musiało najpierw stać się ubogie. Podkreślenia wymaga fakt, że na początku okresu obserwacji gospodarstwo nie mogło być już w trakcie pobytu w sferze ubóstwa, ponieważ w takim przypadku nie znamy momentu wejścia do tej sfery, a tym samym nie jesteśmy w stanie określić jak długo faktycznie to gospodarstwo oczekuje na wyjście z ubóstwa. Oznacza to, że odrzucamy dane lewostronnie ucięte (brak informacji o czasie startu oczekiwania na wystąpienie zdarzenia).

W analizie czasu oczekiwania na pierwsze wyjście ze sfery ubóstwa (wejście do sfery ubóstwa) w skład zbioru danych wchodziły następujące kolumny:
Nr_gosp – numer gospodarstwa domowego,
Czas – czas przeżycia lub czas cenzurowania,
Cenzurowanie – stan cenzurowania przyjmujący wartość 1 dla obserwacji niecenzurowanej oraz wartość 0 dla obserwacji cenzurowanej.

Fragment zbioru danych[4] przedstawiono w tabeli 1.

Tabela 1.

Struktura danych w analizie Kaplana-Meiera

Nr_gosp	Czas	Cenzurowanie
1	2	1
2	4	0
3	1	1

Źródło: opracowanie własne.

Analizę czasów oczekiwania na kolejne wyjścia ze sfery ubóstwa (wejścia do sfery ubóstwa) przeprowadzono w oparciu o zbiór danych składający się z następujących kolumn:
Nr_gosp – numer gospodarstwa domowego; powtórzony dla każdej nowej luki czasowej,
Czas – czas wystąpienia kolejnego zdarzenia lub czas cenzurowania,
Cenzurowanie – stan cenzurowania. Wszystkie obserwacje przyjmowały wartość 1 (obserwacja niecenzurowana) dla każdego gospodarstwa domowego z wyjątkiem ostatniego, które przyjmowało wartość 0 (obserwacja cenzurowana).

[4] Numery gospodarstw domowych w tabelach 1 oraz 2 nie odpowiadają numerom gospodarstw z bazy danych.

W tabeli 2 przedstawiono fragment zbioru danych.

Tabela 2.

Struktura danych dla zdarzeń powtarzających się

Nr_gosp	Czas	Cenzurowanie
1	2	1
1	1	0
2	4	0
3	1	1
3	1	1
3	1	0

Źródło: opracowanie własne.

Funkcje przeżycia oszacowano i zilustrowano graficznie, używając pakietów `survfit` (zob. Therneau, Lumley, 2015) oraz `survrec` (zob. González i inni, 2015) pakietu R (R Development Core Team, 2015). Pierwszy z pakietów stosowano w przypadku estymatora Kaplana-Meiera, natomiast drugi w przypadku estymatorów Peñy-Strawdermana-Hollandera oraz Wanga-Changa.

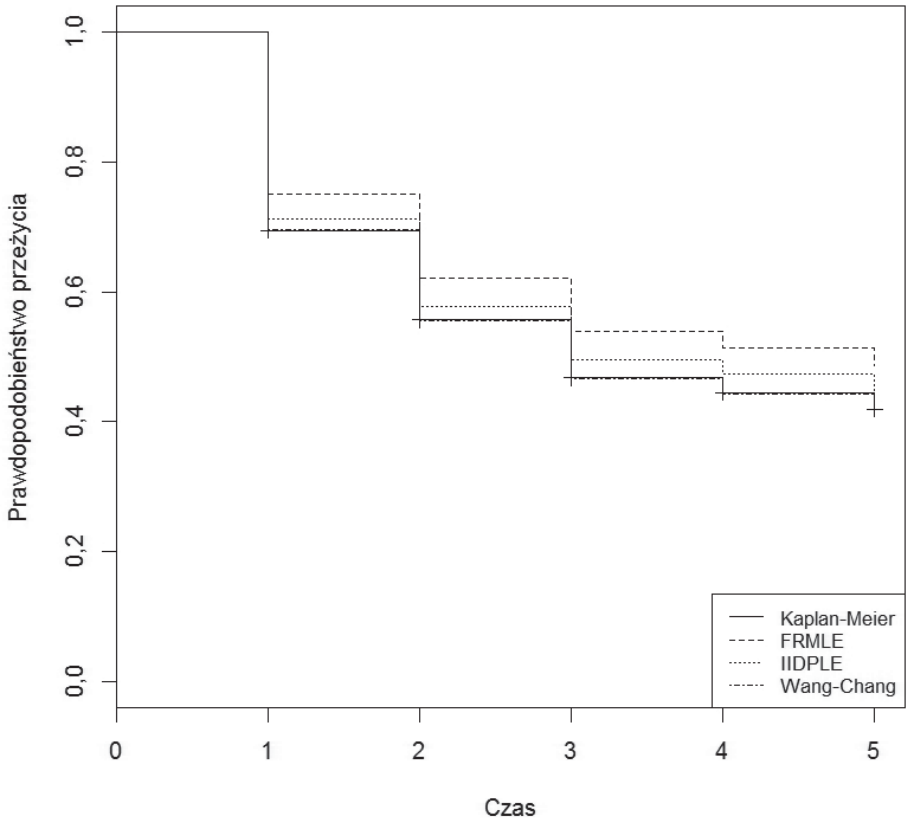
W tabeli 3 oraz na rysunku 6 przedstawiono wyniki oszacowania funkcji przeżycia poza sferą ubóstwa.

Tabela 3.

Funkcje przeżycia poza sferą ubóstwa

Liczba etapów od momentu rozpoczęcia okresu poza ubóstwem	Estymator			
	Kaplana-Meiera	Wanga-Changa	IIDPLE	FRMLE (α = 2755)
1	0,694	0,696	0,712	0,750
2	0,557	0,555	0,578	0,621
3	0,468	0,466	0,495	0,538
4	0,444	0,443	0,473	0,515
5	0,420	0,418	0,447	0,487
Czas przeżycia: mediana średnia	 3 3,16	 3 3,16	 3 3,26	 4 3,42

Źródło: obliczenia własne na podstawie (Rada Monitoringu Społecznego, 2013).



Rysunek 6. Krzywe przeżycia poza sferą ubóstwa
Źródło: jak tabela 3.

Oszacowane funkcje przeżycia wskazują, że ok. 70% gospodarstw domowych pozostaje poza sferą ubóstwa dwa lata lub dłużej (przerwy pomiędzy kolejnymi etapami badania trwały dwa lata), natomiast ponad 40% gospodarstw przeżywa poza sferą ubóstwa dziesięć lat lub dłużej. Najbardziej zbliżone oszacowania uzyskano za pomocą estymatora Wanga-Changa oraz estymatora Kaplana-Meiera analizującego jedynie pierwsze epizody. Jak można zauważyć, pomiędzy oszacowanymi funkcjami przeżycia dla powtarzających się epizodów występują różnice. Z powodu braku formalnych metod statystycznych sprawdzających założenie, że luki czasowe są i.i.d., Peña i inni (2001) zaproponowali, aby rozstrzygnąć tę kwestię stosując metodę graficzną. Uznali, że w przypadku, gdy krzywe przeżycia uzyskane za pomocą IIDPLE, FRMLE oraz Wanga-Changa są zgodne, założenie o niezależności i takich samych rozkładach jest prawdziwe. Na rysunku 6 największe rozbieżności występują pomiędzy oszacowaniami uzyskanymi za pomocą FRMLE a pozostałymi estymatorami, co sugeruje, że założenie i.i.d. nie jest prawdziwe. W praktyce oznacza to, że wnioski dotyczące przeżycia gospodarstw domowych poza sferą ubóstwa należy wyciągać, korzystając z modelu słabości gamma.

Dysponując oszacowanymi i zaprezentowanymi graficznie funkcjami przeżycia, można w łatwy sposób wyznaczyć zarówno medianę (punkt czasu, przy którym wartość funkcji przyjmuje 0,5), jak i średnią (pole pod krzywą przeżycia) czasu przeżycia poza sferą ubóstwa, co pozwoli w pełniejszy sposób zobrazować różnice i podobieństwa pomiędzy uzyskanymi wynikami. Można zauważyć, że obydwie miary przyjmują najwyższe wartości w przypadku FRMLE – średni czas przeżycia gospodarstw domowych poza sferą ubóstwa wynosi niecałe siedem lat oraz połowa gospodarstw przebywa poza sferą ubóstwa co najwyżej osiem lat, a połowa co najmniej osiem lat. Mediany i średnie przyjmują takie same i jednocześnie najniższe wartości w przypadku estymatorów Kaplana-Meiera i Wanga-Changa.

Tabela 4 oraz rysunek 7 przedstawiają wyniki oszacowania funkcji przeżycia w sferze ubóstwa.

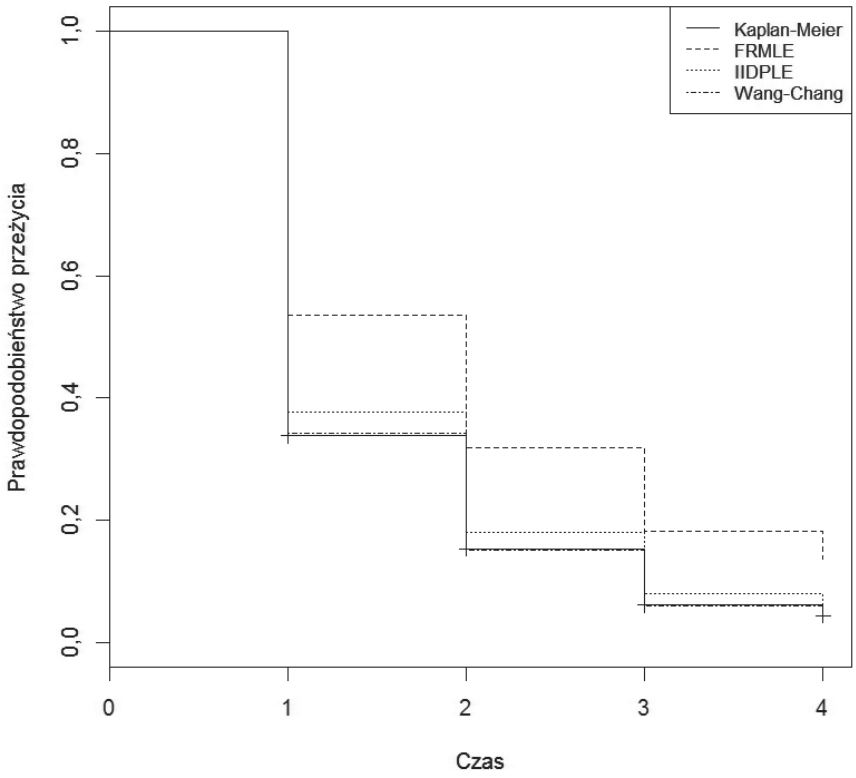
Tabela 4.

Funkcje przeżycia w sferze ubóstwa

Liczba etapów od momentu rozpoczęcia okresu ubóstwa	Estymator			
	Kaplana-Meiera	Wanga-Changa	IIDPLE	FRMLE (α = 2804)
1	0,338	0,342	0,376	0,536
2	0,154	0,152	0,180	0,318
3	0,062	0,061	0,080	0,183
4	0,044	0,043	0,057	0,137
Czas przeżycia: mediana średnia	 1 1,55	 1 1,56	 1 1,64	 2 2,04

Źródło: jak tabela 3.

W przypadku oszacowanych funkcji przeżycia w sferze ubóstwa występują duże różnice pomiędzy estymatorem FRMLE a estymatorami IIDPLE oraz Wanga-Changa (estymator Wanga-Changa ponownie dał rezultaty zbliżone do estymatora Kaplana-Meiera). Na tej podstawie można stwierdzić, że model i.i.d. nie jest odpowiedni w przypadku analizowanych danych. Uzyskane oszacowania za pomocą FRMLE pozwalają stwierdzić, że ponad 50% gospodarstw domowych pozostaje w sferze ubóstwa dwa lata lub dłużej, natomiast niecałe 14% gospodarstw przeżywa w sferze ubóstwa osiem lat lub dłużej.



Rysunek 7. Krzywe przeżycia w sferze ubóstwa
Źródło: jak tabela 3.

Podobnie jak w przypadku przeżycia poza sferą ubóstwa, średnie i mediany czasu przeżycia w sferze ubóstwa przyjęły najwyższe wartości w przypadku FRMLE – gospodarstwa spędzały średnio w sferze ubóstwa cztery lata oraz połowa gospodarstw przeżywała w ubóstwie nie mniej niż cztery lata, a połowa nie więcej niż cztery lata. Średnie i mediany czasu przeżycia przyjęły takie same wartości w przypadku estymatorów Kaplana-Meiera i Wanga-Changa, co potwierdza wcześniej sformułowane wnioski o zbliżonych oszacowaniach uzyskanych za pomocą tych estymatorów.

6. PODSUMOWANIE

W artykule zaprezentowano wyniki analizy czasu trwania gospodarstw domowych w sferze ubóstwa oraz w sferze poza ubóstwem. Wykorzystane w tym celu nieparametryczne estymatory funkcji przeżycia dla zdarzeń powtarzających się pozwoliły stwierdzić, że prawdopodobieństwo pozostania w sferze ubóstwa przez długi czas jest mniejsze niż w przypadku przeżycia poza sferą ubóstwa. Należy podkreślić, że zastosowane estymatory (Wanga-Changa, IIDPLE oraz FRMLE) dały odmienne wyniki. Szczególnie znacząca różnica występuje w przypadku przeżycia w sferze ubóstwa – oszacowania FRMLE są znacznie wyższe niż w przypadku pozostałych estymato-

rów, co sugeruje, że założenie o niezależności i takich samych rozkładach luk czasowych w obrębie gospodarstw domowych nie jest słuszne. Uzyskane wyniki należy jednak traktować ostrożnie. Wejście i wyjście ze sfery ubóstwa może pojawić się w dowolnym momencie, lecz dane panelowe są gromadzone w dyskretnych przedziałach czasu, stąd bardziej odpowiednie byłyby estymatory zakładające czas dyskretny, a nie ciągły jak w przypadku zastosowanych estymatorów. Należy zaznaczyć, że do dnia dzisiejszego nie zaproponowano estymatora dla zdarzeń powtarzających się dla danych o czasie dyskretnym, dlatego estymatory Wanga-Changa, IIDPLE i FRMLE są często stosowane w analizie tego typu danych (np. Gutiérrez i inni, 2011; Hollifield i inni, 2012).

Zgodnie z uwagą poczynioną we wstępie (Allison, 2010), analizę można ograniczyć jedynie do pierwszego epizodu, ponieważ liczba epizodów przypadająca na gospodarstwo domowe była mniejsza niż dwa. Wtedy właściwe są wyniki uzyskane za pomocą estymatora Kaplana-Meiera, które zarówno w przypadku czasu oczekiwania na wyjście, jak i wejście do sfery ubóstwa były najbardziej zbliżone do oszacowań uzyskanych estymatorem Wanga-Changa.

Przedmiotem kolejnych badań będzie analiza czasu trwania gospodarstw w sferze ubóstwa i poza ubóstwem z wykorzystaniem modeli Coxa (zakładających czas ciągły) oraz modeli logitowych (zakładających czas dyskretny). Analiza zostanie przeprowadzona zarówno dla pojedynczych epizodów (czas oczekiwania na pierwsze wejście i wyjście ze sfery ubóstwa), jak i dla powtarzających się zdarzeń.

Anna Sączewska-Piotrowska – Uniwersytet Ekonomiczny w Katowicach

LITERATURA

Allison P. D., (2010), Survival Analysis, w: Hancock G. R., Mueller R. O., (red.), *The Reviewer's Guide to Quantitative Methods in the Social Sciences*, Routlege, New York, 423–424.

Andriopoulou E., Tsakloglou P., (2011), The Determinants of Poverty Transitions in Europe and the Role of Duration Dependence, IZA Discussion Paper No. 5692, Bonn, Germany.

Atkinson T., Cantillon B., Marlier E., Nolan B., (2002), *Social Indicators. The EU and Social Inclusion*, Oxford University Press, New York.

Bane M. J., Ellwood D. T., (1986), Slipping Into and Out of Poverty: The Dynamics of Spells, *The Journal of Human Resources*, 21 (1), 1–23.

Bieszk-Stolorz B., Markowicz I., (2012), *Modele regresji Coxa w analizie bezrobocia*, CeDeWu, Warszawa.

Breslow N., Crowley J., (1974), A Large Sample Study of the Life Table and Product Limit Estimates Under Random Censorship, *Annals of Statistics*, 2 (3), 437–453.

Callens M., Croux C., (2009), Poverty Dynamics in Europe. A Multilevel Discrete-Time Recurrent Hazard Analysis, *International Sociology*, 24 (3), 368–396.

Collett D., (2003), *Modelling Survival Data in Medical Research*, Chapman and Hall/CRC, Boca Raton, Florida.

Fouarge D., Layte R., (2005), Welfare Regimes and Poverty Dynamics: The Duration and Recurrence of Poverty Spells in Europe, *Journal of Social Policy*, 34 (3), 407–426.

Frątczak E., Gach-Ciepiela U., Babiker H., (2005), *Analiza historii zdarzeń. Elementy teorii, wybrane przykłady zastosowań*, SGH, Warszawa.

González J. R., (2006), Inference for a General Class of Models for Recurrent Events with Application to Cancer Data, Universitat Politècnica de Catalunya, Barcelona, praca doktorska.

González J. R., Peña E. A., (2003), Bootstrapping Median Survival with Recurrent Event Data, IX Conferencia Española de Biometría, A Coruña, 2003 May 28–30.

González J. R., Peña E. A., (2004), Estimación no Paramétrica de la Función de Supervivencia Para Datos con Eventos Recurrentes, *Revista Española de Salud Pública*, 78 (2), 189–199.

González J. R., Peña E. A., Strawderman R. L., (2015), Survrec: A Package for Survival Analysis for Recurrent Event Data, URL http://CRAN.R-project.org/package=survrec.

Greenwood M., (1926), The Natural Duration of Cancer, *Reports of Public Health and Related Subjects*, Volume 33, Her Majesty's Stationery Office, London, 1–26.

Gutiérrez E., Lozano S., González J. R., (2011), A Recurrent-Events Survival Analysis of the Duration of Olympic Records, *IMA Journal of Management Mathematics*, 22 (2), 115–128.

Hagenaars A. J. M., de Vos K., (1988), The Definition and Measurement of Poverty, *The Journal of Human Resources*, 23 (2), 211–221.

Hagenaars A. J. M., van Praag B. M. S., (1985), A Synthesis of Poverty Line Definitions, *Journal of the International Association for Research in Income and Wealth*, 31 (2), 139–154.

Hollifield E., Treviño V., Zarn A., (2012), *A Survival Analysis of the Duration of Olympic Records*, arXiv:1207.6133 [stat.AP].

Hosmer D. W., Lemeshow S., May S., (2008), *Applied Survival Analysis. Regression Modeling of Time-to-Event Data*, John Wiley & Sons, Inc., Hoboken, New Jersey.

Kaplan E. L., Meier P., (1958), Nonparametric Estimation From Incomplete Observations, *Journal of the American Statistical Association*, 53 (282), 457–481.

Kelly P. J., Lim L. L.-Y., (2000), Survival Analysis for Recurrent Event Data: An Application to Childhood Infectious Diseases, *Statistics in Medicine*, 19 (1), 13–33.

Kleinbaum D. G., Klein M., (2005), *Survival Analysis. A Self-Learning Text*, Springer, New York.

Layte R., Fouarge D., (2004), The Dynamics of Income Poverty, w: Berthoud R., Iacovou M., (red.), *Social Europe: Living Standards and Welfare States*, Edward Elgar, Cheltenham, 202–224.

Mills M., (2011), *Introducing Survival and Event History Analysis*, SAGE Publications, Los Angeles-London-New Dehli-Singapore-Washington DC.

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

Panek T., Podgórski J., Szulc A., (1999), *Ubóstwo: teoria i praktyka pomiaru*, Monografie i Opracowania 453, SGH, Warszawa.

Peña E. A., Strawderman R. L., Hollander M., (2001), Nonparametric Estimation with Recurrent Data, *Journal of the American Statistical Association*, 96 (456), 1299–1315.

Peterson A. V., (1977), Expressing the Kaplan-Meier Estimator as a Function of Empirical Subsurvival Function, *Journal of the American Statistical Association*, 72 (360), 854–858.

Rada Monitoringu Społecznego, (2013), Diagnoza społeczna 2000–2013: zintegrowana baza danych, http://www.diagnoza.com [29 października 2014 r.].

R Development Core Team, (2015), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, URL http://www.R-project.org.

Rodgers J. R., Rodgers J. L., (1993), Chronic Poverty in the United States, *The Journal of Human Resources*, 28 (1), 25–54.

Stevens A. H., (1994), The Dynamics of Poverty Spells: Updating Bane and Ellwood, *American Economic Review*, 84 (2), 34–37.

Stevens A. H., (1999), Climbing Out of Poverty, Falling Back in. Measuring the Persistence of Poverty Over Multiple Spells, *The Journal of Human Resources*, 34 (3), 557–588.
Therneau T. M., Lumley T., (2015), Survival: A Package for Survival Analysis in S, R Package Version 2.37-7, URL http://CRAN.R-project.org/package=survival.
Topińska I., (2008), Kierunki zmian w statystyce ubóstwa, w: Topińska I., (red.), J. Ciecieląg, A. Szukiełojć-Bieńkuńska, *Pomiar ubóstwa. Zmiany koncepcji i ich znaczenie*, IPiSS, Warszawa, 8–26.
Wang M.-C., Chang S.-H., (1999), Nonparametric Estimation of a Recurrent Survival Function, *Journal of the American Statistical Association*, 94 (445), 146–153.

BADANIE UBÓSTWA Z ZASTOSOWANIEM NIEPARAMETRYCZNEJ ESTYMACJI FUNKCJI PRZEŻYCIA DLA ZDARZEŃ POWTARZAJĄCYCH SIĘ

S t r e s z c z e n i e

W artykule przeprowadzono analizę czasu trwania gospodarstw domowych w sferze ubóstwa oraz w sferze poza ubóstwem. W tym celu wykorzystano estymatory funkcji przeżycia dla zdarzeń powtarzających się: estymator Wanga-Changa oraz dwa estymatory zaproponowane przez Peñę, Strawdermana i Hollandera (IIDPLE oraz FRMLE). Na podstawie uzyskanych wyników można stwierdzić, że prawdopodobieństwo przeżycia gospodarstw domowych w sferze poza ubóstwem przez długi czas jest większe niż w przypadku przeżycia w sferze ubóstwa. Bazując na graficznej metodzie uznano, że najlepszym estymatorem funkcji przeżycia w sferze ubóstwa i w sferze poza ubóstwem jest FRMLE. Oznacza to, że założenie o niezależności i takich samych rozkładach czasów oczekiwania na wystąpienie kolejnych zdarzeń w obrębie gospodarstw domowych nie jest słuszne.

Słowa kluczowe: ubóstwo, funkcja przeżycia, zdarzenia powtarzające się, nieparametryczna estymacja

POVERTY STUDY USING NONPARAMETRIC ESTIMATION OF RECURRENT SURVIVAL FUNCTION

A b s t r a c t

The article analyses households' poverty and nonpoverty duration. For this purpose survival function estimators for recurrent events were used: Wang-Chang estimator and two estimators proposed by Peña, Strawderman and Hollander (IIDPLE and FRMLE). We can conclude that survival probability for a long time out of poverty is greater than in the case of survival in poverty. Based on the graphical method we can conclude that the best estimator of survival in poverty and out of poverty is FRMLE. It means that we cannot assume that interoccurrence times within households are independent and identically distributed.

Keywords: poverty, survival function, recurrent events, nonparametric estimation