PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 106
Ubezpieczenia emerytalne, społeczne i metody aktuarialne 2010

Rafał Korzonek
Uniwersytet Ekonomiczny we Wrocławiu

WYBRANE ESTYMATORY FUNKCJI PRZEŻYCIA

Streszczenie: Ubezpieczenia na życie są oparte na metodach matematycznych i statystycznych, a składki ubezpieczeniowe od dawna oblicza się na podstawie wzorów rachunku prawdopodobieństwa i tablic trwania życia. Pojęciem wymieralności ściśle zajmuje się technika analizy przeżycia, która pierwotnie rozwinęła się w naukach biologicznych i medycznych. Obecnie technika ta, w której główną rolę odgrywa estymacja funkcji przeżycia badanego obiektu, stała się bardzo uniwersalna i coraz częściej jest wykorzystywana w ekonomii i naukach technicznych. W artykule przedstawiono kilka wariantów estymacji tej funkcji, rozważając różne założenia. Jednak w ubezpieczeniach zdrowotnych dane statystyczne to w większości obserwacje ucięte, a najlepszym estymatorem uwzględniającym takie dane jest coraz bardziej popularny estymator Kaplana-Meiera. W artykule poświęcono mu więcej uwagi, przedstawiono kilka jego postaci i sformułowano podstawowe jego własności.

Słowa kluczowe: funkcja przeżycia, estymator Kaplana-Meiera, metoda największej wiarogodności.

1. Wstęp

Technika analizy przeżycia, która pierwotnie rozwinęła się w naukach biologicznych i medycznych, stała się już bardzo uniwersalna i coraz częściej jest wykorzystywana w ekonomii i naukach technicznych. Jednym z przykładów zastosowania tej techniki jest szacowanie długości życia osób ubezpieczonych. Aby móc tego dokonać, należy wyznaczyć funkcję przeżycia badanej grupy osób. W artykule dokonano przeglądu wybranych estymatorów tej funkcji, ze szczególnym uwzględnieniem estymatora Kaplana-Meiera, który ma zastosowanie przy uciętych danych.

2. Estymacja funkcji przeżycia

Ubezpieczenia na życie są oparte na metodach matematycznych i statystycznych, a składki od dawna oblicza się na podstawie wzorów rachunku prawdopodobieństwa i tablic trwania życia, z którymi ściśle łączy się pojęcie funkcji przeżycia. Aby przybliżyć definicję funkcji przeżycia, wprowadźmy potrzebne oznaczenia: niech T_i, $i = 1, 2, \ldots, n$, będą czasami przeżycia, czyli nieujemnymi i niezależnymi zmiennymi losowymi o jednakowym rozkładzie określonym przez dystrybuantę $F(t) = P(T_1 \le t)$. Funkcję

$$S(t) = 1 - F(t) = P(T_1 > t)$$

nazywamy funkcją przeżycia. A zatem ta funkcja informuje o prawdopodobieństwie zdarzenia, że obserwowany obiekt dożyje (przetrwa) do pewnego czasu $t > 0$. Przy założeniu istnienia gęstości $f(t) = F'(t)$ funkcję przeżycia możemy przedstawić w postaci: $S(t) = \int_t^{\infty} f(x)\,dx$.

Warto zwrócić uwagę na to, że jeśli znamy rozkład zmiennych losowych T_i, wówczas znana jest także analityczna postać funkcji $S(t)$ i model przeżycia nazywamy modelem parametrycznym, jeśli natomiast nie znamy rozkładu zmiennych T_i, wówczas mamy do czynienia z modelem nieparametrycznym. Należy również pamiętać, że mając daną funkcję przeżycia, możemy obliczać inne charakterystyki dla zmiennych T_i, jak np.: średni dalszy czas życia, wariancję czy kwantyle.

Okazuje się jednak, że nie zawsze znamy analityczną postać funkcji przeżycia. Pojawia się zatem problem jej estymacji. W artykule przedstawiono kilka podstawowych wariantów estymacji funkcji przeżycia. Wszystkie opisane poniżej estymatory (por. [Johnson, Johnson 1999]) są budowane przy odpowiednich założeniach. Wykorzystywanie ich w praktyce nie jest zbyt powszechne, gdyż wymagane jest spełnienie wszystkich założeń dotyczących rozkładów czasu śmierci i wycofania, jedynie tzw. estymator aktuarialny często pojawia się w naukach aktuarialnych i jest wykorzystywany przy tworzeniu tablic trwania życia.

Przypuśćmy, że oś czasu jest podzielona na rozłączne przedziały postaci $I_j = [t_{j-1}, t_j)$, $j = 1, 2, ...,$ oraz $t_0 = 0$. W każdym przedziale zdefiniujmy następujące stany liczbowe:

d_j – liczba osób zmarłych w przedziale I_j,

w_j – liczba osób, które wycofują się z badań w przedziale I_j,

c_j – liczba osób, które deklarują wycofanie się w przedziale I_j,

d'_j – liczba osób martwych spośród tych, którzy deklarowali wycofanie się,

R_j – liczba osób obecnych na początku przedziału I_j, czyli tzw. grupa ryzyka.

Kilka prostych przekształceń pozwala przedstawić funkcję przeżycia następująco:

$$P(T_1 \geq t_i) = S(t_i) = \prod_{j=1}^{i} P(T_1 \geq t_j \mid T_1 \geq t_{j-1}) = \prod_{j=1}^{i} p_j = \prod_{j=1}^{i} (1 - q_j),$$

gdzie p_j i q_j reprezentują odpowiednio prawdopodobieństwo przeżycia i śmierci podczas przedziału I_j, przy założeniu, że było się żywym na początku przedziału (w chwili t_{j-1}^{-}). Zauważmy, że każdy estymator prawdopodobieństwa śmierci q_j wyznacza estymator funkcji przeżycia S. Zajmijmy się zatem estymacją funkcji przeżycia.

Estymator 1.1

Rozważmy estymator oparty na dystrybuancie empirycznej. Załóżmy, że znamy dokładne czasy śmierci $t_1 < t_2 < \ldots < t_n$. Wówczas naturalnym estymatorem funkcji przeżycia S wydaje się być estymator oparty na dystrybuancie empirycznej:

$$\hat{S}(t) = \begin{cases} 1 & \text{dla } t < t_1, \\ \dfrac{n-i}{n} & \text{dla } t_i \le t < t_{i+1}, \\ 0 & \text{dla } t \ge t_n. \end{cases}$$

z własności dystrybuanty empirycznej zaś wynika, że:

$$E\hat{S}(t) = S(t) \text{ oraz } \operatorname{Var}\hat{S}(t) = \frac{S(t)\left(1-S(t)\right)}{n} = \frac{S(t)F(t)}{n}.$$

Estymator ten, chociaż zdaje się być najprostszy dla funkcji przeżycia, nie sprawdza się w sytuacji, gdy pojawiają się obserwacje ucięte.

Estymator 1.2

Rozważmy tzw. estymator oparty na próbie zredukowanej. Załóżmy, że znamy następujące stany liczbowe: d_j, w_j, R_j. Prawdopodobieństwo śmierci q_j estymujemy przez proporcję osób martwych w przedziale I_j, natomiast nie bierzemy pod uwagę cenzur. A zatem niech

$$\hat{q}_j = \frac{d_j}{R_j - w_j}.$$

Naturalnym estymatorem funkcji przeżycia S jest więc

$$\hat{S}(t) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{R_j - w_j}\right).$$

Należy zauważyć, że również ten estymator nie wykorzystuje informacji niesionych przez obserwacje ucięte.

Estymator 1.3

Rozważamy popularny tzw. estymator aktuarialny. Załóżmy, że znamy następujące stany liczbowe: d_j, w_j, R_j, nie znamy natomiast dokładnych momentów wycofania się osób z badania. W badaniach aktuarialnych przyjmuje się, że wszyscy wycofują się w połowie przedziału I_j. Wówczas estymator prawdopodobieństwa śmierci q_j pozwala otrzymać estymator funkcji przeżycia S postaci:

$$\hat{S}(t)=\prod_{j:t_j<t}\left(1-\frac{d_j}{R_j-\frac{1}{2}w_j}\right).$$

Estymator 1.4

Rozważamy prostą modyfikację poprzedniego przypadku. Załóżmy, jak poprzednio, że znamy stany liczbowe: d_j, w_j, R_j. Teraz zakładamy, że czas śmierci ma rozkład jednostajny na przedziale I_j. Wówczas estymator prawdopodobieństwa śmierci q_j wyznaczony metodą największej wiarogodności pozwala otrzymać estymator funkcji przeżycia S postaci:

$$\hat{S}(t)=\prod_{j:t_j<t}\left(1-\frac{2R_j+d_j-w_j-\sqrt{\left(2R_j+d_j-w_j\right)^2-8R_jd_j}}{2R_j}\right).$$

Estymator 1.5

Załóżmy teraz, że znamy stany liczbowe: d_j, w_j, c_j, d'_j, R_j. Ponadto zakładamy, że czas śmierci ma rozkład wykładniczy na przedziale I_j, a czasy wycofania mają rozkład jednostajny na przedziale I_j. Wówczas estymator prawdopodobieństwa śmierci q_j wyznaczony metodą największej wiarogodności pozwala otrzymać estymator funkcji przeżycia S postaci:

$$\hat{S}(t)=\prod_{j:t_j<t}\left(\frac{-d'_j+\sqrt{d'^2_j+4\left(2R_j-c_j\right)\left(2R_{j+1}+w_j\right)}}{2\left(2R_j-c_j\right)}\right)^2.$$

3. Estymator Kaplana-Meiera

Przedstawione powyżej estymatory są wykorzystywane w ściśle określonych sytuacjach, gdy spełnione są wszystkie wymagane założenia. Okazuje się, że w analizie danych dotyczących przeżycia częstym problemem jest utrata informacji o dokładnym czasie przeżycia. Takie dane są nazywane danymi uciętymi (niekompletnymi). Powyższe estymatory albo nie biorą tych danych pod uwagę, albo nie przypisują im wytarczającej wagi.

W technice analizy przeżycia pojawiającym się najczęściej estymatorem funkcji przeżycia jest estymator Kaplana-Meiera. O jego popularności świadczy choćby fakt, że znalazł się on w podstawowych pakietach obliczeniowych, np. w programie Statistica w module analizy przeżycia. W tym programie ma on następującą postać

$$S(t)=\prod_{j=1}^{t}\left(\frac{n-j}{n-j+1}\right)^{\delta(j)}.$$

W tym wzorze n oznacza liczbę obserwowanych przypadków, $\delta(j)$ oznacza status obserwacji, który wynosi 1, jeśli j-ty przypadek nie jest ucięty (kompletny), lub 0, jeśli jest ucięty, natomiast iloczyn przebiega po wszystkich przypadkach mniejszych od t. Powyższy estymator nazywa się także estymatorem limitu iloczynowego. Po raz pierwszy został wprowadzony przez E. Kaplana i P. Meiera [1958].

Na podstawie analizy powyższego wzoru nietrudno zauważyć, że wykres estymatora Kaplana-Meiera jest funkcją schodkową, dokładniej mówiąc składa się z poziomych odcinków schodzących do zera. Zwiększanie próby powoduje powstawanie większej liczby coraz krótszych odcinków. W granicy proces $\{S(t), t\geq 0\}$ dąży do prawdziwej funkcji przeżycia.

Można się zastanowić, dlaczego ten estymator jest taki ważny. Otóż oprócz tego, że nie odrzuca on obserwacji uciętych poprzez przypisanie im etykiety braku danych, ale bierze je pod uwagę przy obliczaniu funkcji przeżycia, przewaga metody Kaplana-Meiera nad innymi metodami polega także na tym, że na uzyskiwane oceny nie wpływa grupowanie danych w przedziałach czasowych.

3.1. Postać estymatora

Niech, jak poprzednio, T_i będą czasami przeżycia, a C_i będą chwilami obcinającymi (czyli nieujemne i niezależne zmienne losowe o jednakowym rozkładzie). Rozważmy elementy losowe (tzw. model losowego obcinania):

$$(Z_i, \delta_i),\ \ i=1,2,\ldots,n,$$

gdzie: $Z_i = T_i \wedge C_i$ oraz $\delta_i = 1_{(0,C_i)}(T_i)$.

Niech $S(t)=P(T_j>t)$, $F(t)=1-S(t)$, $\pi_j(t)=P(Z_j\geq t)$.

Zdefiniujmy teraz statystyki pozycyjne $Z_{1:n}\leq Z_{2:n}\leq\ldots\leq Z_{n:n}$ dla zmiennych losowych $Z_1, Z_2, \ldots, Z_n$ i niech $I_j=\left[Z_{j-1:n}, Z_{j:n}\right)$ będą przedziałami czasowymi.

Dla każdego t definiujemy grupę ryzyka $R(t)$ jako liczbę podmiotów żywych w chwili t. Niech $R(t)=\sum_{j=1}^{n}1_{\{Z_j\geq t\}}$ oraz niech $M(t)$ będzie liczbą martwych zaobserwowanych w chwili t, czyli $M(t)=\sum_{j=1}^{n}\delta_j 1_{\{Z_j=t\}}$.

Naturalny estymator prawdopodobieństwa śmierci q_j powinien wykorzystywać liczbę martwych w przedziale I_j i grupę ryzyka na początku tego przedziału.

Definicja 3.1. Estymator Kaplana-Meiera

Dla każdego *t* estymatorem $\hat{S}_{KM}$ funkcji przeżycia *S* jest

$$\hat{S}_{KM}(t) = \prod_{\{Z_i \le t\}} \left(1 - \frac{M(Z_i)}{R(Z_i)}\right).$$

Można przyjąć, że w praktyce nie mamy do czynienia z równymi wartościami obserwacji. Jeśli więc w jednej chwili nie nastąpi więcej niż jedno zdarzenie, to $M(Z_i)$ przyjmuje wartość δ_i. Ponadto, w tym przypadku, $R(Z_i) = n - i + 1$. Zatem estymator Kaplana-Meiera możemy zapisać w postaci:

$$\hat{S}_{KM}(t) = \prod_{i:Z_{i:n} \le t} \left(1 - \frac{\delta_{i:n}}{n - i + 1}\right).$$

Wariancja estymatora wyraża się wówczas wzorem:

$$Var\,\hat{S}_{KM}(t) = \hat{S}_{KM}{}^2(t) \sum_{Z_i < t} \frac{\delta_i}{(n-i)(n-i+1)}.$$

Okazuje się także, iż estymator Kaplana-Meiera można wyznaczyć za pomocą metody największej wiarygodności empirycznej.

4. Własności estymatora Kaplana-Meiera

W tym punkcie zajmiemy się własnościami estymatora Kaplana-Meiera. Metody matematyczne, których tutaj użyto, mimo niezbyt prostej postaci estymatora, pozwoliły szybciej i prościej sformułować i udowodnić podstawowe jego własności. Dzięki tym metodom scharakteryzowano obciążenie estymatora Kaplana-Meiera, pokazano, że jego obciążenie szybko znika, wprowadzono estymator na wariancję oraz wykazano jednostajną zgodność w prawdopodobieństwie estymatora Kaplana-Meiera.

Zachowując oznaczenia z poprzednich punktów artykułu, rozważmy następujące procesy stochastyczne:

$$\bar{N}(t) = \sum_{j=1}^{n} N_j(t) = \sum_{j=1}^{n} 1_{\{Z_j \le t,\, \delta_j = 1\}},$$

$$\bar{Y}(t) = \sum_{j=1}^{n} Y_j(t) = \sum_{j=1}^{n} 1_{\{Z_j \ge t\}},$$

oraz skumulowaną funkcję hazardu daną wzorem

$$\Lambda(t) = \int_0^t \left(1 - F(s-)\right)^{-1} dF(s).$$

Rozważamy następującą naturalną filtrację $\{F_t : t \geq 0\}$ daną przez:

$$F_t = \sigma\left\{N_j(s), N_j^c(s) : 0 \leq s \leq t,\ j = 1, \ldots, n\right\},\ \ t \geq 0.$$

4.1. Postać estymatora

Przejdziemy do estymacji funkcji przeżycia. Metodę estymacji dostarcza nam relacja pomiędzy Λ i S. Z definicji mamy:

$$\Lambda(t) = \int_0^t (1 - F(s-))^{-1} dF(s),\ \text{zatem}\ d\Lambda(s) = \frac{dF(s)}{1 - F(s-)}.$$

Po dokonaniu prostych przekształceń otrzymujemy równanie rekurencyjne, które po rozwiązaniu daje estymator Kaplana-Meiera postaci

$$\hat{S}(t) = \prod_{s \leq t}\left(1 - \frac{\Delta\bar{N}(s)}{\bar{Y}(s)}\right),$$

4.2. Obciążenie estymatora

Przytoczymy *prosty lemat mówiący o obciążeniu estymatora Kaplana-Meiera.*

Lemat 4.1

Jeśli $S(t) > 0$, oraz $T = \inf\{s : \bar{Y}(s) = 0\}$, to mamy:

(1) $E\left(\hat{S}(t) - S(t)\right) = EB(t) = E\left(1_{\{T<t\}} \frac{\hat{S}(T)\left(S(T) - S(t)\right)}{S(T)}\right) \geq 0,$

(2) jeśli ponadto $\pi_j(s) = \pi(s)$ dla każdego j, to $E\left(\hat{S}(t) - S(t)\right) \leq \left(1 - S(t)\right)\left(1 - \pi(t)\right)^n$.

Z powyższego lematu wynika, że estymator Kaplana-Meiera jest obciążony tylko wtedy, gdy jest dodatnie prawdopodobieństwo zdarzenia, że $\hat{S}(T) > 0$ i $S(t) < S(T)$.

4.3. Asymptotyczna normalność estymatora

Zauważmy, że estymator Kaplana-Meiera $\hat{S}$ jest asymptotycznie normalny $N\left(0, \sigma^2(t)\right)$, a wyrażenie na $\sigma^2(t)$ można przedstawić w prostej postaci:

$$\sigma^2(t) = -S^2(t)\int_0^t \frac{S^2(s-)}{S^2(s)}\left(\pi(s)\right)^{-1} \frac{S(s)}{S^2(s-)} dS(s) = -S^2(t)\int_0^t \frac{dS(s)}{\pi(s)S(s)}.$$

4.4. Zgodność estymatora Kaplana-Meiera

W [Fleming, Harrington 2005] możemy znaleźć twierdzenie mówiące o zgodności estymatora Kaplana-Meiera.

Twierdzenie 4.1

Niech T będzie czasem życia, to znaczy zmienną losową o dystrybuancie $F(s)=P\{T\le s\}$ i skumulowanej funkcji hazardu postaci $\Lambda(s)=\int_0^s \frac{dF(v)}{1-F(v)}$.

Jeśli $t\in(0,\infty]$ jest takie, że $\bar{Y}(t)\xrightarrow{P}\infty$ gdy $n\to\infty$, to

$\sup_{0\le s\le t}\left|\int_0^s \frac{d\bar{N}(v)}{\bar{Y}(v)}-\Lambda(s)\right|\xrightarrow{P}0$ gdy $n\to\infty$ oraz $\sup_{0\le s\le t}\left|\hat{F}(s)-F(s)\right|\xrightarrow{P}0$ gdy $n\to\infty$, gdzie $1-\hat{F}$ jest estymatorem Kaplana-Meiera.

Twierdzenie to pokazuje, że estymator Kaplana-Meiera $\hat{S}(t)$ jest jednostajnie zgodnym estymatorem funkcji przeżycia $S(t)$.

5. Przykład zastosowania estymatora Kaplana-Meiera

Rozważmy grupę 20 osób ubezpieczonych na życie, które w trakcie okresu ubezpieczenia przebyły zawał serca. W wyniku symulacji uzyskano obserwacje zawarte w poniższej tabeli. Czas obserwacji liczony jest w miesiącach od daty zakończenia leczenia klinicznego, status równy 1 oznacza zgon, a status równy 0 oznacza, że osoba żyła w dniu zakończenia obserwacji.

Tabela 1. Czas obserwacji

Lp.	Czas	Status	Lp.	Czas	Status	Lp.	Czas	Status	Lp.	Czas	Status
1	2,367	1	6	6,677	0	11	8,378	0	16	15,639	0
2	2,399	1	7	7,197	1	12	9,495	1	17	15,704	1
3	2,784	0	8	8,016	0	13	10,567	0	18	19,701	0
4	3,189	1	9	8,131	1	14	11,677	1	19	21,955	1
5	3,929	1	10	8,317	0	15	11,765	0	20	24,309	0

Źródło: opracowanie własne.

Jak już wcześniej wspomniano, estymator przedstawia funkcję schodkową, wystarczy więc określić jego wartość w punktach nieciągłości. Estymator ten daje nam następującą funkcję przeżycia (część opisu funkcji pomijamy).

$t_0=0$, $S(t_0)=P(Y>t_0)=1$

$t_1=2{,}367$ $S(t_1)=P(Y>t_1)=0{,}95$

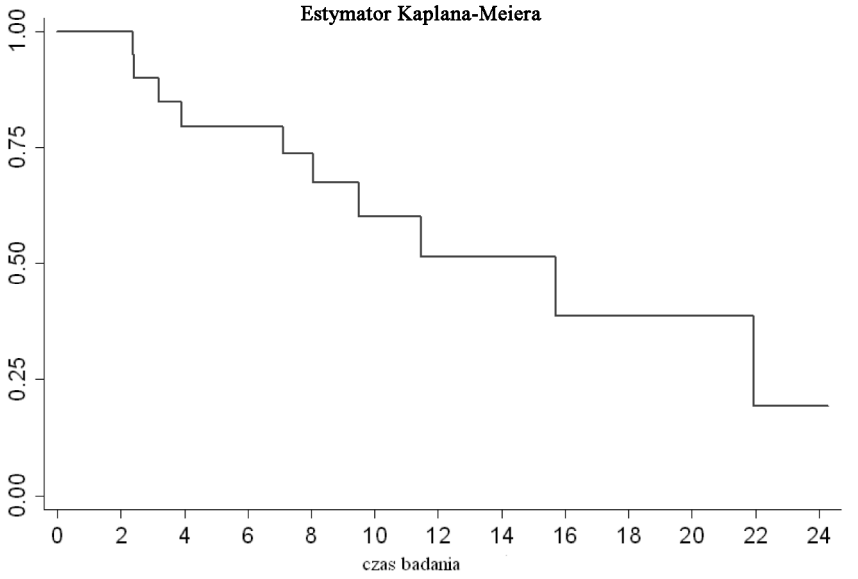
$t_2 = 2{,}399$ $\quad S(t_2) = P(Y > t_2) = 0{,}90$

….

$t_{18} = 19{,}701$ $\quad S(t_{18}) = P(Y > t_{18}) = 0{,}386$

$t_{19} = 21{,}955$ $\quad S(t_{19}) = P(Y > t_{19}) = 0{,}193$

$t_{20} = 24{,}309$ $\quad S(t_{20}) = P(Y > t_{20}) = 0{,}193$



Rys. 1. Wykres estymatora Kaplana-Meiera

Źródło: opracowanie własne.

Na podstawie wyestymowanej w ten sposób funkcji przeżycia ubezpieczyciel może dokonywać oceny w zakresie ryzyka ubezpieczeniowego, a także szacować dalszy czas życia osoby ubezpieczonej.

Literatura

Badania statystyczne w ubezpieczeniach, red. J. Hozer, Szczecin 2002.

Droesbeke J., Fichet B., *Analyse s tatistique des durées de vie*, „Economica" 1989.

Fleming T., Harrington D., *Counting Processes and Survival Analysis*, Wiley 2005.

Johnson R., Johnson N., *Survival Models and Data Analysis*, New York 1999.

Kaplan E., Meier P., *Nonparametric estimation from incomplete observations*, "Journal of American Statistical Association" 1958, no. 53.

CHOSEN ESTIMATORS OF SURVIVAL FUNCTION

Summary: Life insurance has already been based on mathematical and statistical methods for a few centuries, and insurance premiums have been calculated on the grounds of the probability calculus and life tables. The survival analysis technique, originally developed in biological and medical sciences, deals with the death rate. This technique has already become very widespread and is more and more often used in economics and technical theories. In each of these cases there appears a need to estimate the survival function of the studied object. There are a few variants of the estimation of this function in the article. Each of the described estimators is built on different assumptions. However, the statistical data in health insurance are largely truncated observations and the best estimator considering such data is the most popular Kaplan-Meier's estimator. The article devotes most attention to this estimator and presents a few of its modifications as well as its basic properties.