ZESZYTY NAUKOWE Instytutu Zarządzania i Marketingu Akademii im. Jana Długosza w Częstochowie
Seria: Pragmata tes Oikonomias 2006, z. I

Stanisława Ostasiewicz

# Analiza trwania życia w populacjach niejednorodnych

## Wstęp

Współczesne badania wykazały, że nasz genom jest zaprogramowany na 120 lat życia. Laboratoryjnie zbadano, że pojedyncze organy mogą funkcjonować nawet 140 lat. Mimo iż w ciągu ostatnich 100 000 lat nie odnotowano żadnych zmian genomu, średnia długość życia ludzkiego znacznie się zwiększyła w ciągu minionych 100 lat. Określenia typu starość, podeszły wiek wypierane są określeniem bardziej neutralnym długowieczność. Gdyby ludzie umierali tylko ze starości, to żyliby przynajmniej 100 lat. Zjawisko to zainteresowało głównie demografów i biologów. Chcieli bowiem wiedzieć, co decyduje i co wpływa na długość życia ludzkiego oprócz wieku biologicznego. Przesunięcie się granicy długości życia ludzkiego wpłynęło też na zmianę stosowanej terminologii. Przyczyny wcześniejszych zgonów stanowią od dawna przedmiot zainteresowania zarówno biologów, jak i demografów. Ostatnio problematyką długowieczności i zróżnicowaniem czasu życia ludzkiego zainteresowali się statystycy. Próbując wyjaśnić ten problem, wykorzystali metodologię nauk probabilistycznych, polegającą na modelowania zjawisk witalności i umieralności.

Akceptując tę metodologię, należy przyjąć, że długość trwania życia osoby w wieku $x$ lat scharakteryzowanej za pomocą pewnych obserwowalnych cech $Y_1, Y_2, ..., Y_n$, oraz nieobserwowalnej bezpośrednio ukrytej cechy $Z$ jest wielkością losową, którą oznaczymy symbolem $T$.

Wielkość $T$, tak jak każdą skalarną zmienną losową, można scharakteryzować za pomocą takich funkcji, jak dystrybuanta, gęstość, funkcja przeżycia, funkcja intensywności, a także za pomocą pewnych parametrów liczbowych, takich jak oczekiwana długość życia, dalsze trwanie życia itp.

Dystrybuantę zmiennej losowej $T$ oznaczamy jako $F_T(t|x,y,z)$ i zgodnie ze standardową jej definicją przyjmujemy, że

$$F_T(t|x,y,z) = P(T \leq t|x,y,z),$$

gdzie $x$ jest to wiek osoby, $y=(y_1, y_2,..., y_n)$ jest to wektor zaobserwowanych wartości cech $Y_1, Y_2,..., Y_n$, zaś $Z$ oznacza nieobserwowalną wielkość cechy ukrytej, nazywanej słabowitością (ang. *frailty*).

Modelowanie i analiza czasu trwania życia ludzkiego polega na sformułowaniu modelu, tzn. określeniu wyrażenia definiującego dystrybuantę lub inną równoważną jej funkcję, identyfikacji, estymacji, i weryfikacji tego modelu. Rozpatrzmy trzy grupy modeli, które symbolicznie przedstawimy następująco:

1) $F_T(t|x)$,
2) $F_T(t|x, y)$,
3) $F_T(t|x, y, z)$.

# 1. Tradycyjne modele trwania życia

Przyjmijmy, że zamiast zapisu $F_T(t|x)$ stosowany będzie tradycyjny zapis aktuarialny tzn. $F_x(t)$, przy czym gdy $x=0$, stosuje się jeszcze bardziej uproszczony zapis $F(t)$.

Wielkość $F(t)$ oznacza prawdopodobieństwo, że osoba dopiero co urodzona, tzn. osoba w wieku 0 lat, nie przeżyje do momentu t, $t \geq 0$, czyli $F(t) = P(T_0 \leq T)$. Za pomocą dystrybuanty $F(t)$ można określić dystrybuantę czasu życia osoby w dowolnym wieku $x$ według następującego wzoru:

$$F_x(t) = \frac{F(x+t) - F(x)}{1 - F(x)}.$$

Wyrażenie występujące w mianowniku tego wzoru nazywa się funkcją przeżycia osoby w wieku 0 lat. W przypadku dowolnego wieku $x$, funkcję przeżycia określa się następująco:

$$S_T(t|x) = 1 - F_T(t|x).$$

Zarówno na podstawie dystrybuanty $F_T(t|x)$, jak i funkcji przeżycia $S_T(t|x)$ definiowana jest inna funkcja charakteryzująca czas życia ludzkiego, a mianowicie funkcja intensywności umieralności.

Kilka równoważnych definicji tej funkcji podano niżej (por. [5, 9, 10]).

$$\mu(t|x) = -\frac{1}{S_T(t|x)} \frac{dS_T(t|x)}{dt}$$

$$\mu(t|x) = \frac{1}{1 - F_x(t)} \frac{dF_x(t)}{dt}$$

$$\mu(t|x) = \frac{f_x(t)}{1 - F_x(t)}$$

gdzie $f_x(t)$ jest to pochodna funkcji $F_x(t)$ względem zmiennej t.

W przypadku gdy $x = 0$, zamiast $\mu(t|0)$ stosuje się zapis $\mu(t)$.

Konkretną postać funkcji $\mu(t)$ po raz pierwszy zaproponował A. De Moivre w 1729 r:

$$\mu(t) = (\omega - t)^{-1},$$

gdzie $\omega$ jest to górna granica wieku życia ludzkiego (np. 100 lub 120 lat).

W 1825 roku Gompertz zaproponował następującą postać funkcji intensywności

$$\mu(t) = b \cdot c^t,$$

gdzie $b$ oraz $c$ są to parametry, które ustala się w zależności od populacji.

Parametr $b$ w tej funkcji określa początkową stopę umieralności i traktowany jest jako niezależny od wieku, zaś parametr $c$ często definiowany jest w postaci $\exp(w)$, wówczas tak zwane prawo Gompertza zapisuje się następująco:

$$\mu(t) = b \cdot \exp(w \cdot t).$$

Nieco ogólniejszą postać tego prawa podał Makeham w 1860 roku:

$$\mu(t) = a + b \cdot c^t.$$

Każde z podanych praw jest funkcją czasu $t$ tzn. określa intensywność umieralności spowodowanej wyłącznie czasem, czyli to, co możemy określić jako umieralność *biologiczną* ze starości.

Powszechnie wiadomo jednak, że na umieralność mają wpływ różne czynniki *środowiskowe*, a także pewne *cechy ukryte* poszczególnych osób. Ze względu na to, że cechy środowiskowe mają zupełnie inną naturę aniżeli cechy ukryte posiadane przez indywidualne osoby istnieją dwa różne podejścia do uwzględniania tych cech w modelach umieralności.

## 2. Modele regresyjne

W przypadku cech środowiskowych, które można bezpośrednio obserwować i mierzyć, wyróżnia się dwa typy modeli:

— modele przyśpieszonego ryzyka,

— modele proporcjonalnego ryzyka.

Istota modeli przyśpieszonego ryzyka polega na tym, że modyfikacji intensywności umieralności dokonuje się poprzez przeskalowanie czasu. Stąd też wynika nazwa tych modeli: czas życia jednostki może być formalnie przyśpieszony.

Modele tego typu mają więc następującą postać:

$$\mu(t, x) \equiv \mu(t) = \mu_0(t; g(x)),$$

gdzie $g(x)$ jest odpowiednio zdefiniowaną funkcją skalowania czasu, która może być zależna od cech środowiskowych.

O wiele bardziej popularne i mające znacznie większe zastosowanie zarówno w demografii, jak i medycynie są modele Coxa.

Intuicyjnie idea tych modeli polega na tym, że podstawową lub bazową funkcję intensywności umieralności, oznaczaną jako $\mu_0(t)$, modyfikuje się, mnożąc ją przez pewien czynnik proporcjonalności zależny od obserwowalnych cech środowiskowych. To znaczy skalowany jest w tym przypadku nie czas, lecz intensywność umieralności. Ten rodzaj modeli prezentowany jest w następnym paragrafie tego artykułu.

Załóżmy, że intensywność umieralności w określonej populacji zależy od momentu czasu $t$, w którym ta intensywność jest określona, od wieku osoby $x$ oraz od pewnych obserwowalnych cech $Y_1, Y_2, \ldots, Y_p$ charakteryzujących tę osobę. Zaobserwowane wartości tych cech traktowane będą dalej jako wektor liczbowy $y = (y_1, y_2, \ldots, y_p)$. Załóżmy ponadto, że intensywność umieralności dla całej populacji bez uwzględniania cech objaśniających oznaczona będzie symbolem $\mu_0(t|x)$ lub równoważnie w postaci $\mu_0(t; x)$.

Ponieważ wszystkie funkcje trwania życia zawsze zależą od wieku osoby, który zwykle traktowany jest jako parametr, a nie jako argument funkcji. Stąd też często, w celu uproszczenia zapisów, jest on pomijany. Jeżeli więc nie będzie to prowadziło do nieporozumień, to zamiast $\mu(t|x)$ stosowany będzie krótszy zapis $\mu(t)$.

W 1972 roku sir D. Cox zaproponował słynny dziś model, nazwany jego imieniem, który określa zależność trwania życia ludzkiego od cech objaśniających $Y_1, Y_2, ..., Y_p$.

Ogólna postać tego modelu jest następująca (por. [4]):

$$\frac{\mu(t;x,y)dt}{1-\mu(t;x,y)dt} = \exp(\beta' y)\frac{\mu_0(t;x,y)dt}{1-\mu_0(t;x,y)dt}.$$

Warto podkreślić, że ogólność tego modelu polega na tym, że obejmuje on przypadki, gdy czas traktowany jest jako wielkość ciągła oraz gdy jest on rozpatrywany tylko dyskretnie. Ta ważna cecha ogólności tego modelu wynika z zastosowania tu uniwersalnej symboliki matematycznej.

Zauważmy, że zapis $\mu(t)dt$ oznacza następujące prawdopodobieństwo:

$$\mu(t)dt = P(t \le T \le t + dt),$$

które różnie traktujemy w przypadku czasu ciągłego i czasu dyskretnego.

Jeżeli zmienną losową $T$ traktujemy jako zmienną losową typu ciągłego, wówczas lewa strona powyższego wzoru ma postać:

$$\frac{\mu(t;x,y)dt}{1-\mu(t;x,y)dt} = \mu(t;x,y)dt.$$

Model Coxa przyjmuje wówczas następującą postać:

$$\mu(t;x,y) = \mu_0(t;x)\cdot\exp(\beta' y).$$

Pomijając parametr wieku osoby $x$, mamy:

$$\mu(t;y) = \mu_0(t)\cdot\exp(\beta' y).$$

Taka postać modelu Coxa nazywa się modelem proporcjonalnego ryzyka, gdyż funkcję intensywności umieralności interpretuje się jako ryzyko umieralności. Ryzyko to, w przypadku uwzględnienia zewnętrznych czynników w postaci cech $Y_1, Y_2, ..., Y_p$, jest proporcjonalne do ryzyka zależnego tylko od czasu $t$. Współczynnikiem proporcjonalności jest wielkość $\exp(\beta' y)$.

W przypadku dyskretnej wersji modelu Coxa mamy więc do czynienia z tzw. modelem proporcjonalnych szans. Wynika to z tego, że wielkość $\mu/(1-\mu)$ nazywana jest szansą (ang. *odds*).

Jeżeli zaś czas życia osoby traktowany jest w sposób dyskretny, to model Coxa przyjmuje następującą postać:

$$\frac{\mu(t_i,y)}{1-\mu(t_i,y)dt} = \frac{\mu_0(t_i)}{1-\mu_0(t_i)}\exp(\beta' y).$$

Wprowadzając oznaczenie $\alpha_i = \mu_0(t_i)$, wówczas dyskretny wariant modelu Coxa zapisać można następująco (por. [1, 4]):

$$\mu(t_i,y) = \frac{\exp(\alpha_i + \beta' y)}{1+\exp(\alpha_i + \beta' y)}$$

i widzimy, że jest to model logistyczny.

W obu przypadkach model Coxa określają nieznane parametry $\beta_1, \beta_2, ..., \beta_p$, które trzeba estymować na podstawie danych empirycznych.

W celu estymacji parametrów $\beta_1, \beta_2, ..., \beta_p$ przyjmujemy, że dokonano obserwacji czasu życia oraz indywidualnych charakterystyk n osób. Przyjmijmy, że momenty czasu w których następowały zgony, są następujące:

$$t_{(1)} < t_{(2)} < ... < t_{(n)},$$

przy czym $m_i$ jest to liczba zgonów, jakie nastąpiły do momentu $t_{(i)}$.

Symbol $y_{(i)}$ oznacza wektor wartości cech (charakterystyk) tej osoby, która zmarła przed upływem czasu $t_{(i)}$. Symbolem $R(t)$ oznaczamy zaś zbiór tych osób, które do momentu $t$ włącznie były obecne w badanej populacji, tzn. ani nie opuściły populacji, ani też nie umarły.

Wykorzystując te oznaczenia przedstawimy teraz funkcje wiarygodności (por. [3]):

$$L(\beta) = \prod_{j=1}^{k} \frac{\exp(\beta' y_{(j)})}{\sum_{i \in R_j} \exp(\beta' y_i)},$$

gdzie $R_j = R(t_{(j)})$.

Ze względu na dość skomplikowaną symbolikę użytą w tym wyrażeniu rozpatrzymy prosty przykład (por. [3]).

Załóżmy, że obserwowana populacja składa się z 5 osób, oznaczonych jako A, B, C, D, E.

Załóżmy, że osoba C umiera po upływie 1 roku od rozpoczęcia obserwacji, osoba E wyjeżdża za granicę po upływie 2 lat od rozpoczęcia obserwacji, osoba A umiera po 3 latach, osoba B wyjeżdża po 3,5 latach, zaś osoba D umiera po 4 latach.

Tak więc w tym przypadku mamy trzy momenty, w których następują zgony:

$$t_{(1)} = 1,\ t_{(2)} = 3,\ t_3 = 4.$$

$R(t_{(1)}) = \{A, B, C, D, E\}$ – do momentu $t_{(1)}$ włącznie wszystkie osoby były obserwowane,

$R(t_{(2)}) = \{A, B, D\}$ – do momentu $t_{(2)}$ nie było już osób C i E,

$R(t_{(3)}) = \{D\}$.

Funkcja wiarogodności jest więc w tym przypadku następująca:

$$L(\beta) = \frac{\exp(\beta' y_c)}{\exp(\beta' y_A) + \exp(\beta' y_B) + \exp(\beta' y_c) + \exp(\beta' y_D) + \exp(\beta' y_E)} \times$$

$$\frac{\exp(\beta' y_c)}{\exp(\beta' y_A) + \exp(\beta' y_B) + \exp(\beta' y_D)} \times \frac{\exp(\beta' y_D)}{\exp(\beta' y_D)}$$

W celu wyznaczenia wartości estymatorów parametrów $\beta_1, \beta_2, ..., \beta_p$ należy rozwiązać układ równań:

$$\frac{\partial \log L(\beta)}{\partial \beta} = 0.$$

Układ taki rozwiązuje się iteracyjnie.

Cox wykazał, że uzyskane w ten sposób estymatory mają asymptotyczny rozkład normalny:

$$\hat{\beta} \sim N(\beta; I^{-1}(\beta)),$$

gdzie

$$I(\beta) = -\frac{\partial^2}{\partial \beta^2} \log L(\beta).$$

Na podstawie wyznaczonych estymatorów $\hat{\beta}$ można określić funkcję intensywności oraz wszystkie inne funkcje z nią związane, na przykład funkcję przeżycia.

Funkcję intensywności estymuje się następująco (por. [3, 4]):

$$\mu(t, y) = \hat{\mu}_0(t)\exp(\hat{\beta}'y).$$

W celu uzyskania występującego w tym wyrażeniu estymatora bazowej funkcji intensywności $\mu_0(t)$ przyjmuje się założenie, że jej niezerowe wartości są tylko w punktach zaobserwowanych zgonów. Estymację wartości $\mu_0(t_{ij})$ uzyskuje się wówczas następująco (por. [3, 7]):

$$\hat{\mu}_0(t_{ij}) = 1 - \hat{\xi}_j,$$

gdzie $\hat{\xi}_j$ jest to rozwiązanie następującego równania:

$$\sum_{i \in D_j} \frac{\exp(\hat{\beta}'y_i)}{1 - \hat{\xi}_j \exp(\hat{\beta}'y_i)} = \sum_{i \in R_j} \exp(\hat{\beta}'y_i),$$

gdzie $D_j$ jest to zbiór wszystkich osób, które zmarły w momencie $t_{(j)}$, zaś $R_j$ jest to zbiór osób narażonych na ryzyko, określony tak jak poprzednio.

Jeśli w każdym momencie $t_{(j)}$ był tylko jeden zgon, to rozwiązanie równania jest następujące:

$$\hat{\xi}_j = \left(1 - \frac{\exp(\hat{\beta}'y_{(j)})}{\sum_{i \in R_j} \exp(\hat{\beta}'y_{(i)})}\right) \exp(-\hat{\beta}'y_{(j)}).$$

Przyjmując kolejne założenie, że intensywność umieralności jest stała między dwoma punktami $t_{(j)}$ oraz $t_{(j+1)}$, łatwo jest określić bazową funkcję przeżycia (por. [3]):

$$\hat{S}_0(t) = \prod_{j=1}^{r} \hat{\xi}_j \quad \text{dla} \quad t_{(r)} \le t \le t_{(r+1)}, \quad r = 1,2,...,k\text{-}1.$$

Na podstawie bazowej funkcji intensywności można określić funkcję intensywności zależną od cech objaśniających.

## 3. Modelowanie słabowitości

Załóżmy tak jak poprzednio, że $T_x$ to zmienna losowa oznaczająca przyszły czas życia osoby, która dożyła wieku $x$ lat. Przyjmijmy teraz, że ten czas życia zależy od pewnej nieobserwowalnej cechy ukrytej, którą traktuje się jako zmienną losową $Z$, a jej dystrybuantę oznacza się symbolem F(z). Rozkład zmiennej losowej $T_x$ zależy więc teraz od dwóch parametrów: wieku osoby oraz jej słabowitości. Pierwszy parametr, zgodnie z tradycją aktuarialną, oznaczymy symbolem $x$, zaś słabowitość – symbolem $z$. Dystrybuantę zmiennej losowej $T_x$ oznaczymy teraz w postaci $F_x(t,z)$ lub jako $F(t,x,z)$. Natomiast funkcję gęstości jako $f(t,x,z)$ lub jako $f_x(t,z)$. W celu uprosz-

czenia zapisów wiek $x$ jest często pomijany. Zamiast $F_x(t,z)$ lub $f_x(t,z)$ zapisuje się krócej w postaci $f(t,z)$ oraz $F(t,z)$.

Bezwarunkowy, czyli brzegowy, rozkład trwania życia określa się jako następującą mieszankę rozkładów:

$$f(t) = \int f(t,z)dF(z).$$

Bez przyjęcia dodatkowych założeń rozkład taki nie jest, niestety, identyfikowalny.

Najprostsze założenie przyjmowane w celu umożliwienia identyfikacji jest takie, że zmienna nieobserwowalna $Z$ ma warunkową funkcję intensywności

$$\mu(t,1) \equiv \mu(t|Z=1) = \frac{f(t|Z=1)}{1-f(t|Z=1)}$$

wpływa w sposób multyplikatywny. To znaczy funkcja intensywności określona jest następująco:

$$\mu(t,Z) = Z \cdot \mu(t,1).$$

Przyjmując konwencję, że $\mu(t) \equiv \mu(t,1)$ otrzymujemy:

$$\mu(t,Z) = Z \cdot \mu(t)$$

Zależność ta nazywa się modelem słabowitości.

Model ten niezależnie od siebie sformułowali różni autorzy, R.E. Beard w 1959 roku, T. Lancaster w 1979 roku, w tym samym roku model ten zaprezentowali J.W. Vaupel, K.G. Manton oraz E. Stallard (por. [2]). Następnie był on modyfikowany i uogólniany na wiele różnych sposób (por. [1, 6, 7]).

Rozpatrzmy model najprostszy, a mianowicie model proporcjonalnego wpływu czynnika słabowitości na czas trwania życia konkretnej osoby w porównaniu z ogólną umieralnością w populacji. Korzystając z warunkowej funkcji intensywności umieralności $\mu(t|z)$, określimy rozkład czasu trwania życia za pomocą warunkowej funkcji przeżycia w sposób następujący:

$$S(t|z) = \exp(-\int_0^t \mu(u|z)du.$$

Ponieważ przyjmujemy, że $\mu(t|Z) = Z \cdot \mu_0(t)$, to bezwarunkową funkcję przeżycia określimy następująco:

$$S(t) = \int S(t|z)f_z(z)dz = \int \exp(-zH_0(t))f_z(z)dz,$$

gdzie $H_0(t) = \int_0^t \mu_0(s)ds$.

Przyjmując, że $Z \sim \Gamma(k,\lambda)$,

czyli

$$f_z(z) = \frac{\lambda^k}{\Gamma(k)} z^{k-1}e^{-\lambda z}, \quad \text{dla} \quad z > 0,$$

otrzymujemy

$$S(t) = \left[1 + \frac{H_0(t)}{\lambda}\right]^{-k}.$$

Jeżeli dodatkowo przyjmiemy, że $E(Z)=1$, zaś $V(Z)=\sigma^2$, to funkcja przeżycia w całej populacji ma następującą postać:

$$S(t)=\left[1+\sigma^2 H_0(t)\right]^{-1/\sigma^2}.$$

Natomiast funkcja intensywności umieralności w tym przypadku ma postać (por. [1]):

$$\mu(t)=\left[1+\sigma^2 H_0(t)\right]^{-1}\cdot\mu_0(t)$$

Czyli

$$\mu(t)=\mu_0(t)\cdot S(t)^{\sigma^2}.$$

Gęstość rozkładu zmiennej losowej $Z$ dla osób, które przeżyły okres $t$ $f_z(z|T>t)$, określić teraz można następująco:

$$f_z(z|T>t)=\frac{S(t)\cdot f_z(x)}{S(t)}.$$

Po podstawieniach otrzymujemy:

$$f_z(z|T>t)=\frac{[\lambda+\mathrm{H}_0(t)]^{\mathrm{k}}}{\Gamma(k)}z^{k-1}e^{-(\lambda-\mathrm{H}_0(t))z}.$$

Jest to, jak widać, rozkład gamma z takim samym parametrem kształtu $k$, ale parametr skali jest teraz równy wielkości $\lambda+H_0(t)$.

W pracy [6] wykazano, że własność taką ma rodzina rozkładów wykładniczych.

## Literatura


1. Barbi E., *Assessing the rate of ageing of the human population*, Max-Planck-Institute for Demographic Research, Working Paper WP 2003-008, March 2003.
2. Butt Z., Haberman S., *Application of frailty-based mortality models using generalized linear models*, „ASTIN Bulletin", vol. 34, No. 1, 2004, 175–197.
3. Collet D., *Modelling survival data in medical research*, Chapman & Hall, London, 1994.
4. Cox D. R., *Regression models and life tables*, „Journal of the Royal Stat. Soc." Ser. B. 34. 1972, 187–202.
5. Doan O. (red.), *Ubezpieczenia życiowe*, Poltex, Warszawa 1995.
6. Hougart P., *Life table methods for heteregeneous populations*, „Biometrika", 71, 1984, 75–83.
7. Lancaster T., *Econometric methods for the duration of unemployment*, „Econometrica", 47, 1979, 939–956.
8. Manton K.G., Stallard E., Vaupel J., *Alternative models for the hetrerogenity of mortality risks among the aged*, „Journal of the American Stat. Association", 81, 1986, 635–644.
9. Ostasiewicz S. (red.), *Metody oceny i porządkowania ryzyka w ubezpieczeniach życiowych*, AE Wrocław, Wrocław 2000.
10. Ostasiewicz S., *Składki w wybranych typach ubezpieczeń życiowych*, AE Wrocław, Wrocław 2000.

11. Thather A. R., *The long-term pattern of adult mortality and the highest attained age*, „Journal of the Royal Statist. Soc. A", 162, part 1, 1999, 5–30.
12. Vaupel J., Manton K.G., Stallard E., *The impact of heterogeneity in individual frailty on the dynamics of mortality,* „Demography", volume 16, number 3, August 1979.