PRZEGLĄD STATYSTYCZNY
R. LVI – ZESZYT 2 – 2009

IWONA MARKOWICZ, BEATA STOLORZ

MODEL PROPORCJONALNEGO HAZARDU COXA PRZY RÓŻNYCH SPOSOBACH KODOWANIA ZMIENNYCH

1. WSTĘP

Metody analizy przeżycia są coraz częściej stosowane w badaniach zjawisk społeczno-ekonomicznych[1]. Ze względu na brak konieczności znajomości rozkładu badanej zmiennej losowej szczególną wagę przywiązuje się do modeli nieparametrycznych bądź semiparametrycznych. Coraz powszechniej wykorzystywane są one do badania zjawisk innych niż czas trwania życia ludzkiego. Przeglądu metodologii analizy historii zdarzeń i ich aplikacji do badania czasu funkcjonowania firm autorki artykułu dokonały w ramach realizacji grantu MNiSW (N 111 011 31/1109). Warunkiem stosowania modeli analizy przeżycia jest odpowiednia baza danych umożliwiająca wyznaczenie czasu trwania zdefiniowanego stanu dla poszczególnych jednostek badanej zbiorowości. Zazwyczaj są to badania retrospektywne z wykorzystaniem sporządzanych rejestrów. Przykładem takiej bazy danych jest rejestr bezrobotnych.

Celem artykułu jest wskazanie wpływu sposobu kodowania zmiennych na oszacowania parametrów modelu regresji Coxa i ich interpretację. Autorki przedstawiły również związek między parametrami modelu szacowanymi dla danych zakodowanych w dwojaki sposób. Badaną kohortę stanowią osoby bezrobotne wyrejestrowane w określonym okresie czasu. Podziału na podgrupy dokonano ze względu na wiek, który jest determinantą czasu poszukiwania pracy, co autorki wykazały we wcześniejszych badaniach [7].

2. DANE STATYSTYCZNE WYKORZYSTANE W ANALIZIE

Analiza czasu oczekiwania na pracę została przeprowadzona w oparciu o indywidualne dane o bezrobotnych wyrejestrowanych z Powiatowego Urzędu Pracy w Szczecinie w I kwartale 2007 roku. Uzyskane informacje dotyczyły wieku i powodu wyrejestrowania. Powody wyrejestrowania były różne, natomiast powód, który został uznany za zdarzenie kończące obserwację to podjęcie przez dotychczasowego bezrobotnego pracy. Osoby wyrejestrowane z innych przyczyn, takich jak podjecie nauki, wyjazd za granicę, odmowa przyjęcia propozycji zatrudnienie, niestawienie się w PUP w wyznaczonym terminie, czy osiągnięcie wieku emerytalnego, stanowią obserwacje cenzurowane. Dla

[1] Por. [1], [6], [7].

tej grupy nie można ustalić okresu oczekiwania na pracę. Analizie poddano ogółem 4237 osób. Kategorie wieku zostały pogrupowane według klasyfikacji stosowanej przez PUP. Spośród wszystkich wyrejestrowanych dla 46% osób powodem było znalezienie pracy. Strukturę badanej zbiorowości przedstawiono w tabeli 1.

Tabela 1

Charakterystyka ilościowa badanych bezrobotnych

Cecha			Obserwacje		Razem
			pełne	cenzurowane	
Wiek	⟨18, 25)	(1)	431	569	1000
	⟨25, 35)	(2)	779	719	1498
	⟨35, 45)	(3)	310	369	679
	⟨45, 55)	(4)	361	474	835
	⟨55, 60)	(5)	61	112	173
	⟨60, 65)	(6)	4	48	52
Ogółem			1946	2291	4237

Źródło: opracowanie własne.

3. MODEL PROPORCJONALNEGO HAZARDU COXA[2]

Do zbadania wpływu potencjalnej zmiennej na czas pozostawania w rejestrze bezrobotnych nie można zastosować modeli regresji wielorakiej ze względu na nieznajomość rozkładu zmiennej zależnej oraz występowanie obserwacji cenzurowanych. Model proporcjonalnego hazardu Coxa zakłada, że funkcja hazardu [4] jest funkcją zmiennych niezależnych, którą można zapisać następująco [1]:

$$h(t{:}x_1, x_2, \ldots, x_n) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n) \tag{1}$$

gdzie:

$h(t{:}\ x_1, x_2, \ldots, x_n)$ – wynikowy hazard (szansa) przy danych n zmiennych niezależnych $x_1, x_2, \ldots, x_n$ i odpowiednim czasie przeżycia (oczekiwania),

$h_0(t)$ – hazard (szansa) odniesienia lub zerowa linia hazardu,

$\beta_1, \beta_2, \ldots, \beta_n$ – współczynniki modelu,

t – czas obserwacji.

Bazowa wartość $h_0(t)$ hazardu jest tą wartością hazardu, dla której wszystkie zmienne niezależne są równe zero.

[2] Por. [3].

Wieloczynnikowy model proporcjonalnego hazardu Coxa umożliwia ocenę jednoczesnego wpływu wielu zmiennych na czas do wystąpienia określonego zdarzenia.

4. SPOSOBY KODOWANIA ZMIENNYCH

Ze względu na różne sposoby kodowania zmiennych można obliczyć różne rodzaje ryzyka względnego. W artykule zostaną przedstawione dwa z nich, zgodnie z procedurami przedstawionymi przez Hosmer i Lemeshow [5].

Pierwszy rodzaj kodowania umożliwia wyznaczenie szansy opuszczenia przez bezrobotnego rejestru PUP w stosunku do wybranej kategorii danej zmiennej. Sposób kodowania w przypadku, gdy do porównania wybrano pierwszą grupę wieku przedstawia tabela 2. Kodowanie cech umożliwiło zastąpienie cechy ilościowej (wiek w latach) cechą kategoryzowaną (kodowanie 0-1). Poszczególne przedziały wieku ponumerowano od 1 do 6. Ponieważ jako odniesienie przyjęto pierwszy przedział wieku ($\langle 18, 25)$), to cechy w modelu oznaczono jako $Wiek\,(i,\ 1)$ dla $i \in \{2, 3, 4, 5, 6\}$.

Tabela 2

I rodzaj kodowania

Wiek bezrobotnych	Cecha *Wiek* (*i*, 1)				
	Wiek (2,1)	*Wiek* (3,1)	*Wiek* (4,1)	*Wiek* (5,1)	*Wiek* (6,1)
$\langle 18, 25)$	0	0	0	0	0
$\langle 25, 35)$	1	0	0	0	0
$\langle 35, 45)$	0	1	0	0	0
$\langle 45, 55)$	0	0	1	0	0
$\langle 55, 60)$	0	0	0	1	0
$\langle 60, 65)$	0	0	0	0	1

Źródło: opracowanie własne.

Wyniki estymacji parametrów modelu Coxa przedstawiono w tabeli 3. Jest to model ze zmiennymi kategoryzowanymi $Wiek\,(i,\ 1)$ dla $i \in \{2, 3, 4, 5, 6\}$.

Tabela 3

Wyniki estymacji parametrów modelu Coxa przy zastosowaniu kodowania I

	Cecha	β_i	Błąd parametru	Wartość statystyki t	$\exp(\beta_i)$	Statystyka Walda	p
Kodowanie I	*Wiek* (2,1)	–0,05824	0,060162	–0,96807	0,943423	0,93715	0,333018
	Wiek (3,1)	–0,54685	0,075151	–7,27666	0,578770	52,94979	0,000000
	Wiek (4,1)	–0,66539	0,072208	–9,21489	0,514073	84,91415	0,000000
	Wiek (5,1)	–0,91716	0,137533	–6,66870	0,399651	44,47152	0,000000
	Wiek (6,1)	–2,67375	0,502618	–5,31965	0,068993	28,29865	0,000000

Źródło: obliczenia własne z wykorzystaniem programu *Statistica*.

Oszacowany, metodą częściowej wiarygodności[3], model można przedstawić w następującej postaci:

$$h(t,x_2,x_3,x_4,x_5,x_6) = \\ = h_0(t)\exp(-0{,}05824x_2 - 0{,}54685x_3 - 0{,}66539x_4 - 0{,}91716x_5 - 2{,}67375x_6), \tag{2}$$

gdzie:

$$x_i = \mathit{Wiek}\,(i,\ 1),\ \text{dla}\ i = 2,\ \ldots,\ 6.$$

Wyrażenie $\exp\beta_i$ wyraża w tym przypadku stosunek szansy na znalezienie pracy przez bezrobotnego z i-tej grupy wieku w porównaniu z grupą pierwszą. Przyjmuje się więc, że $\beta_1 = 0$.

Istnieje również możliwość obliczenia szansy względnej między dowolnymi dwiema grupami wieku. Wartość odpowiedniego parametru beta wyznacza się jako stosunek funkcji proporcjonalnego hazardu dla porównywanych kategorii danej zmiennej, przy założeniu stałości pozostałych zmiennych objaśniających[4]. Zmiany ryzyka w zależności od grupy wieku wyznaczono na podstawie wzoru:

$$\mathit{Wiek}(i,j) = \frac{\mathit{Wiek}(i,1)}{\mathit{Wiek}(j,1)} = \frac{\exp\beta_i}{\exp\beta_j} = \exp(\beta_i - \beta_j),\ \text{dla}\ i,j = 2,\ldots,6 \tag{3}$$

a wyniki zaprezentowano w tabeli 4.

[3] Por. [8], s. 29-30, [5], s. 11-14.
[4] Por. [5], s. 123-124.

Tabela 4

Szansa względna uzyskania zatrudnienia wyznaczona na podstawie wzoru (3)

Szansa względna uzyskania zatrudnienia przez bezrobotnych	w stosunku do grupy wieku				
z grupy wieku	⟨18, 25)	⟨25, 35)	⟨35, 45)	⟨45, 55)	⟨55, 60)
⟨25, 35)	0,943423				
⟨35, 45)	0,57877	0,613479			
⟨45, 55)	0,514073	0,544902	0,888216		
⟨55, 60)	0,399651	0,423618	0,690518	0,777421	
⟨60, 65)	0,068993	0,073131	0,119206	0,134209	0,172633

Źródło: obliczenia własne z wykorzystaniem programu *Statistica*.

Drugi rodzaj kodowania umożliwia wyznaczenie szansy opuszczenia rejestru PUP przez bezrobotnego z danej grupy wieku względem średniej całej badanej kohorty (tabela 5). Ponieważ jako odniesienie przyjęto średnią dla kohorty oznaczoną jako s, to cechy w modelu oznaczono jako *Wiek* (i, s) dla $i \in \{1, 2, 3, 4, 5, 6\}$.

Tabela 5

II rodzaj kodowania

Wiek bezrobotnych	Cecha *Wiek* (i, s)				
	Wiek $(2, s)$	*Wiek* $(3, s)$	*Wiek* $(3, s)$	*Wiek* $(4, s)$	*Wiek* $(5, s)$
⟨18, 25)	–1	–1	–1	–1	–1
⟨25, 35)	1	0	0	0	0
⟨35, 45)	0	1	0	0	0
⟨45, 55)	0	0	1	0	0
⟨55, 60)	0	0	0	1	0
⟨60, 65)	0	0	0	0	1

Źródło: opracowanie własne.

W przypadku podziału kohorty na n grup otrzymujemy $n - 1$ estymatorów parametrów $\beta_2^*, \beta_3^*, \ldots, \beta_n^*$, przy czym zachodzi warunek:

$$\sum_{i=1}^{n} \beta_i^* = 0, \tag{4}$$

czyli:

$$\beta_1^* = -\sum_{i=2}^{n} \beta_i^*. \tag{5}$$

Wyniki estymacji parametrów modelu regresji Coxa przedstawiono w tabeli 6.

Tabela 6

Wyniki estymacji parametrów modelu Coxa przy zastosowaniu kodowania II

	Cecha	β_i^*	Błąd parametru	Wartość statystyki t	$\exp(\beta_i^*)$	Statystyka Walda	p
Kodowanie 2	*Wiek* (1, *s*)	0,810257	0,096426	8,402876	2,248487	70,60831	0,000000
	Wiek (2, *s*)	0,75202	0,092535	8,12681	2,121274	66,04506	0,000000
	Wiek (3, *s*)	0,26340	0,099121	2,65740	1,301352	7,06177	0,007879
	Wiek (4, *s*)	0,14486	0,097587	1,48445	1,155882	2,20358	0,137700
	Wiek (5, *s*)	–0,10691	0,136522	–0,78311	0,898605	0,61326	0,433568
	Wiek (6, *s*)	–1,86363	0,417675	–4,46191	0,155109	19,90862	0,000008

Źródło: obliczenia własne z wykorzystaniem programu *Statistica*.

W tym przypadku model ze zmiennymi kategoryzowanymi *Wiek* (*i*, *s*) dla $i \in \{1, 2, 3, 4, 5, 6\}$ ma postać:

$$\begin{gathered} h(t, x_2, x_3, x_4, x_5, x_6) = \\ = h_0(t)\exp(0{,}810257x_1 + 0{,}75202x_2 + 0{,}2634x_3 + 0{,}14486x_4 - 0{,}10697x_5 - 1{,}86363x_6), \end{gathered} \tag{6}$$

gdzie:

$$x_i = \textit{Wiek}\,(i,\ s),\ \text{dla}\ i = 1,\ \ldots,\ 6.$$

Wyrażenie $\exp(\beta_i^*)$ wyraża w tym przypadku stosunek szansy na znalezienie pracy przez bezrobotnego z *i*-tej grupy wieku w porównaniu ze średnią całej kohorty.

Również w przypadku drugiego kodowania można obliczyć szansę względną między *i*-tą grupą wieku, a grupą pierwszą. Można ją wyznaczyć korzystając z zależności [2]:

$$\textit{Wiek}(i,1) = \frac{\exp\beta_i^*}{\exp\beta_1^*} = \exp(\beta_2^* + \ldots 2\beta_i^* + \ldots + \beta_n^*). \tag{7}$$

Analogicznie wyznacza się szansę względną między dowolnymi dwiema grupami wieku:

$$\textit{Wiek}(i,j) = \frac{\textit{Wiek}(i,s)}{\textit{Wiek}(j,s)} = \frac{\exp\beta_i^*}{\exp\beta_j^*} = \frac{\exp\beta_i}{\exp\beta_j}. \tag{8}$$

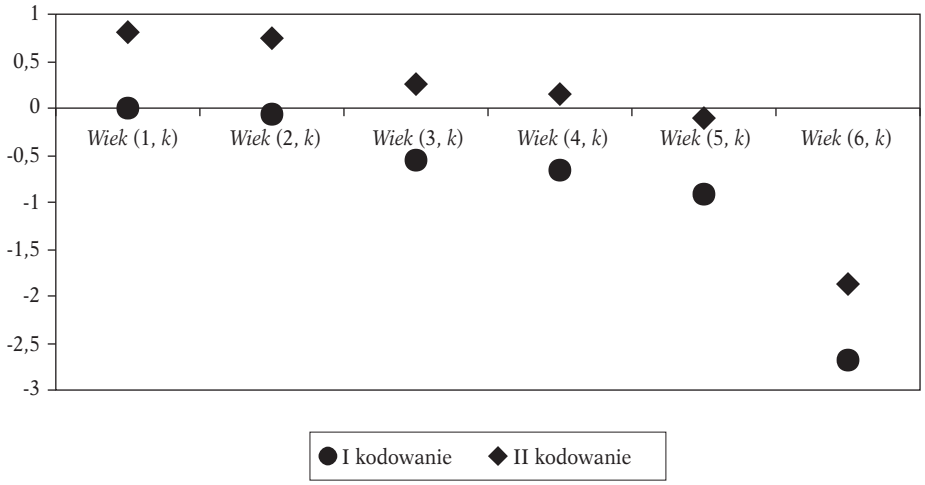
Otrzymane wartości szansy względnej są takie same, jak w przypadku kodowania pierwszego (tabela 4).

Stopień dopasowania modelu przy zastosowaniu obu sposobów kodowania jest oczywiście taki sam, wartość statystyki χ^2 wynosi 229,844 przy poziomie istotności $p = 0{,}0000$.

5. ZWIĄZKI MIĘDZY PARAMETRAMI MODELU COXA PRZY ZASTOSOWANIU OBU SPOSOBÓW KODOWANIA

Zastosowanie poszczególnych sposobów kodowania daje w wyniku różne oszacowania parametrów modelu Coxa i inna jest też ich interpretacja. Kodowanie I pozwala na wyznaczenie szansy na znalezienie pracy przez bezrobotnego z danej grupy wieku (i) względem możliwości zdobycia zatrudnienia osób z grupy pierwszej. Przykładowo bezrobotni w wieku od 45 do 55 lat mają prawie o połowę mniejszą szansę podjęcia pracy niż osoby w wieku od 18 do 25 lat. Przy interpretacji parametrów, w przypadku zastosowania kodowania II, punktem odniesienia jest średnia szansa znalezienia pracy całej badanej zbiorowości. Bezrobotni w wieku od 45 do 55 lat w tym przypadku o 15% szybciej znajdowali zatrudnienie niż przeciętnie w całej kohorcie.

Jak już wskazano oszacowania parametrów, jak też ich interpretacja, przy zastosowaniu obu sposobów kodowania są różne, ale przedstawiając na wykresie (rysunek 1) wyznaczone wartości parametrów modelu regresji Coxa można zauważyć istnienie pewnej zależności.



Rysunek 1. Wartości oszacowanych parametrów β i β^* (stała różnica między parametrami)

Źródło: opracowanie własne.

Wartości szansy względnej wyznaczone ze wzoru (3) przy zastosowaniu kodowania I i ze wzoru (8) przy zastosowaniu kodowania II, które zaprezentowano w tabeli 4, są

jednakowe. W związku z tym powinna istnieć zależność między parametrami oznaczonymi w artykule jako β i β^*.

Korzystając z zależności:

$$\exp\left(\beta_2^* + \ldots + 2\beta_i^* + \ldots + \beta_n^*\right) = \exp\beta_i, \quad \text{dla } i = 2, 3, \ldots, n \tag{9}$$

oraz

$$\exp\left(\beta_1^* + \beta_2^* + \ldots + \beta_i^* + \ldots + \beta_n^*\right) = \exp\beta_1 \tag{10}$$

otrzymujemy wzory przejścia między dwoma omówionymi sposobami kodowania:

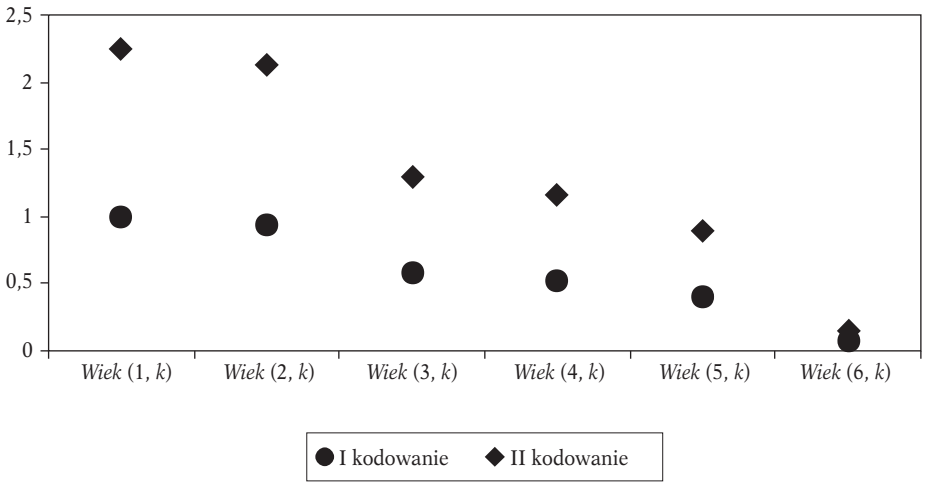
$$\beta_i^* = \beta_i - \frac{1}{n}\sum_{k=1}^{n}\beta_k, \text{ dla } k = 1, 2, \ldots, n. \tag{11}$$

Różnica między parametrami β i β^* równa się:

$$\beta_i - \beta_i^* = \frac{1}{n}\sum_{k=1}^{n}\beta_k, \text{ dla } k = 1, 2, \ldots, n, \tag{12}$$

czyli jest stała i równa się średniej arytmetycznej parametrów β_i, uzyskanych w wyniku pierwszego kodowania.

Ponieważ potwierdzono istnienie zależności między parametrami β i β^* można przypuszczać, że również zachodzi związek między szansami względnymi, wyznaczonymi odpowiednio w stosunku do pierwszej grupy wieku oraz średniego czasu pozostawania bez pracy. Oszacowane wartości szansy względnej w przypadku kodowania I i II przedstawiono na rysunku 2.



Rysunek 2. Oszacowane wartości szansy względnej w przypadku kodowania I i II (stały stosunek szans względnych)

Źródło: opracowanie własne.

Korzystając ze wzoru (12) można znaleźć związek między $\exp(\beta_i)$ i $\exp(\beta_i^*)$:

$$\frac{\exp(\beta_i)}{\exp(\beta_i^*)} = \exp\left(\frac{1}{n}\sum_{k=1}^{n}\beta_k\right), \text{ dla } k = 1, 2, \ldots, n. \tag{13}$$

Ze wzoru (13) można odczytać, że stosunki szansy względnej w przypadku kodowania pierwszego i drugiego są stałe i równe $\exp\left(\frac{1}{n}\sum_{k=1}^{n}\beta_k\right)$.

6. PODSUMOWANIE

Z przedstawionych badań wynikają następujące wnioski:

– parametry modelu Coxa można wyznaczyć stosując dwa sposoby kodowania zmiennych wpływających na czas poszukiwania pracy,

– kodowanie 0-1 oznaczone w artykule jako kodowanie I wymusza określenie podgrupy odniesienia; w analizowanym przykładzie wybrano podgrupę pierwszą, czyli bezrobotnych w wieku od 18 do 25 lat,

– jako grupę odniesienia można przyjąć dowolną podgrupę, którą badacz chce wyróżnić; może to być na przykład grupa najliczniejsza, najstarsza itp.,

– stosując kodowanie −1-0-1, oznaczone w artykule jako kodowanie II, punktem odniesienia jest średnia całej kohorty; w tym przypadku nie ma znaczenia wybór podgrupy, która zostanie oznaczona przez –1,

– między parametrami modelu proporcjonalnego hazardu Coxa w przypadku obu sposobów kodowania zmiennych istnieje związek; różnica między odpowiadającymi sobie parametrami jest stała i równa średniej arytmetycznej z parametrów otrzymanych dla kodowania 0-1;

– wyznaczenie szansy względnej na podjęcie pracy dla dowolnych dwóch podgrup jest możliwe przy zastosowaniu dowolnego rodzaju kodowania.

Uniwersytet Szczeciński

LITERATURA

[1] Bednarski T., [2005], *Ocena przydatności danych Bael dla charakterystyki rozkładu czasu poszukiwania pracy na przykładzie danych z lat 2001-2002*, Studia Ekonomiczne, nr 4, Instytut Nauk Ekonomicznych PAN, Warszawa.

[2] Colett D., [2003], *Modelling Survival Data in Medical Research*, Chapman & Hall/CRC, Boca Raton, Floryda.

[3] Cox D.R., Oakes D., [1984], *Analysis of Survival Data*, Chapman and Hall, London.

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

[5] Hosmer D.W., Lemeshow S., [1999], *Applied Survival Analysis. Regression Modeling of Time to Event Data*, John Wiley & Sons, INC, New York.

[6] Hozer J. (red.), [2002], *Badania statystyczne w ubezpieczeniach*, Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Szczecin.

[7] Markowicz I., Stolorz B., [2007], *Determinants of Labour Seeking Time Resulting From Labour Demand on Szczecin Labour Market*, The labour demand in the modern economy, Economics & Competition Policy, No. 10, Katedra Mikroekonomii Uniwersytetu Szczecińskiego, Szczecin.
[8] Rossa A., [2005], *Metody estymacji rozkładu czasu trwania zjawisk dla danych cenzurowanych oraz ich zastosowania*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Praca wpłynęła do redakcji w listopadzie 2008 r.

MODEL PROPORCJONALNEGO HAZARDU COXA PRZY RÓŻNYCH SPOSOBACH KODOWANIA ZMIENNYCH

Streszczenie

Metody analizy przeżycia są coraz częściej stosowane w badaniach zjawisk społeczno-ekonomicznych. Ze względu na brak konieczności znajomości rozkładu badanej zmiennej losowej szczególną wagę przywiązuje się do modeli nieparametrycznych bądź semiparametrycznych. Coraz powszechniej wykorzystywane są one do badania zjawisk innych niż czas trwania życia ludzkiego. Warunkiem stosowania modeli analizy przeżycia jest odpowiednia baza danych umożliwiająca wyznaczenie czasu trwania zdefiniowanego stanu dla poszczególnych jednostek badanej zbiorowości. Zazwyczaj są to badania retrospektywne z wykorzystaniem sporządzanych rejestrów. Przykładem takiej bazy danych jest rejestr bezrobotnych.

Celem artykułu jest wskazanie wpływu sposobu kodowania zmiennych na oszacowania parametrów modelu regresji Coxa i ich interpretację. Autorki przedstawiły również związek między parametrami modelu szacowanymi dla danych zakodowanych w dwojaki sposób. Badaną kohortę stanowią osoby bezrobotne wyrejestrowane w określonym okresie czasu. Podziału na podgrupy dokonano ze względu na wiek, który jest determinantą czasu poszukiwania pracy.

Słowa kluczowe: analiza przeżycia, modele semiparametryczne, model regresji Coxa, kodowanie.

THE COX PROPORTIONAL HAZARD MODEL FOR DIFFERENT METHODS OF ENCRYPTION OF VARIABLES

Summary

Methods of survival analysis are more and more often used in analysis of social and economic occurrences. Due to lack of distributional information regarding the random variable, much attention is put on non-parametric or semi-parametric models. They are more and more commonly used for analysis of occurrences different than life expectancy. The condition of use of models of survival analysis is appropriate database that makes possible estimation of duration time of defined state for particular elements of analysed population. They are usually retrospective analyses with use of records. The example of such database is unemployment records.

The article presents results of analysis of influence of encryption of variables on estimation of parameters of the Cox proportional hazard model and their interpretation. The authors also presented correlation between parameters of the model estimated for the data encrypted in two ways. The cohort consisted of the unemployed persons unregistered in specific period. Sub-clusters were allocated with respect to age that is a determinant of period of waiting for a job.

Key words: survival analysis, semi-parametric models, Cox regression model, encryption.