DZIAŁ NAUKOWY

MATEMATYKA STOSOWANA TOM 13/54 2011

Agnieszka Deszyńska (Kraków)

Model hazardów proporcjonalnych Coxa

Streszczenie. Celem pracy jest prezentacja modelu hazardów proporcjonalnych Coxa (ang. *Cox proportional hazards model*), charakteryzujących go własności oraz metod estymacji jego parametrów. Znajduje on zastosowanie w analizie przeżycia przy przewidywaniu szans przetrwania pewnych obiektów (najczęściej pacjentów w badaniach medycznych). Istotną zaletą modelu jest możliwość uwzględnienia w nim danych niepełnych, które często pojawiają się w przeprowadzanych badaniach — zarówno w sposób losowy, jak i celowy. Model Coxa sprawdza się szczególnie dobrze w sytuacji, gdy interesujące jest określenie skuteczności sposobu leczenia w sensie porównawczym, czyli w odniesieniu do innych terapii. Terminologia i przykłady zaczerpnięte są na ogół z medycyny, ale opisany model stosuje się również np. w socjologii, kryminalistyce czy inżynierii.
Słowa kluczowe: model Coxa, hazard, analiza przeżycia.

1. Wstęp. Analiza przeżycia jest gałęzią statystyki obejmującą metody badania procesów, w których obiektem zainteresowania jest czas, jaki upłynie do wystąpienia pewnego zdarzenia. Jako takie zajście rozpatrywać możemy np. śmierć pacjenta w badaniach medycznych (stąd nazwa działu statystyki), awarię urządzenia (w inżynierii), popełnienie przestępstwa (w kryminalistyce), rozwód, ukończenie szkoły czy odejście pracownika z firmy. Ze względu na tak dużą różnorodność tych wydarzeń można spotkać się z innym niż w biostatystyce nazewnictwem — przykładowo, w inżynierii stosuje się określenie analiza awarii, zaś w socjologii — analiza historii zdarzeń.

Główne pytania, na jakie analiza przeżycia pomaga udzielić nam odpowiedzi, to:

- Jaka część populacji przetrwa pewien okres czasu?
- Jak długo będą żyć ci, którzy przetrwają?
- Czy należy brać pod uwagę więcej niż jeden czynnik sprzyjający niepowodzeniu?
- Jakie określone okoliczności bądź cechy charakterystyczne badanego obiektu wpływają na szanse przetrwania?

Jedną z typowych metod analizy przeżycia są modele regresyjne, w tym model hazardów proporcjonalnych Coxa [2], na którym skupia się niniejsza praca.

2. Podstawowe definicje.

Definicja 2.1 (*Czas życia*). Czasem życia nazywamy taką zmienną losową T, dla której

$$P(T \geqslant 0) = 1,$$

tzn. przyjmującą z prawdopodobieństwem 1 wartości nieujemne.

W dalszych definicjach F oznaczać będzie dystrybuantę czasu życia T, zaś f — jego gęstość.

Definicja 2.2 (*Funkcja przeżycia*). Funkcja przeżycia jest to funkcja dana wzorem

$$S(x) = P(T \geqslant x) = 1 - F(x^-). \tag{2.1}$$

Zgodnie z powyższą definicją funkcja przeżycia określa prawdopodobieństwo, że badany obiekt będzie żył przez co najmniej x. W przypadku dyskretnego rozkładu prawdopodobieństwa czasu życia funkcja przeżycia wyraża się przez

$$S(x) = \sum_{k:x_k \geqslant x} p_k,$$

gdzie $p_k = P(T = x_k)$, zaś w przypadku rozkładu ciągłego

$$S(x) = \int_x^\infty f(t)dt.$$

Definicja 2.3. (*Funkcja hazardu*). Funkcja hazardu jest to funkcja dana wzorem

$$h(x) = \frac{f(x)}{S(x)}. \tag{2.2}$$

Z definicji funkcji przeżycia wynika, że

$$S'(x) = -F'(x) = -f(x),$$

zatem funkcja przeżycia i funkcja hazardu powiązane są zależnością

$$h(x) = \frac{-S'(x)}{S(x)} = -\frac{d}{dx}\ln S(x). \tag{2.3}$$

Funkcję hazardu definiuje się również jako

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr[(t \leqslant T < t + \Delta t)|(T \geqslant t)]}{\Delta t}. \tag{2.4}$$

Należy zauważyć, że funkcja hazardu nie jest prawdopodobieństwem warunkowym (np. może przyjmować wartości większe od 1), można ją natomiast interpretować jako oczekiwaną liczbę zdarzeń na jednostkę badaną w jednostce czasu.

Definicja 2.4 (*Skumulowana funkcja hazardu*). Skumulowana funkcja hazardu jest to funkcja dana wzorem

$$H(x) = \int_0^x h(u)du. \tag{2.5}$$

Z powyższej definicji oraz zależności (2.3) wynika, że

$$H(x) = -\ln S(x).$$

3. Przykłady funkcji hazardu. Mając daną funkcję hazardu dla danego rozkładu, można — korzystając z zależności podanych w poprzednim rozdziale — łatwo wyznaczyć pozostałe funkcje opisujące ten rozkład, tzn. funkcję gęstości, skumulowaną funkcję hazardu oraz funkcję przeżycia. Kilka przykładów prostych funkcji hazardu (określających użyteczne i stosowane w analizie przeżycia rozkłady) znajduje się poniżej.

- *Funkcja stała.*
  Dla funkcji hazardu określonej wzorem
  $$h(t) = c > 0.$$
  funkcja gęstości przybiera postać
  $$f(t) = ce^{-ct},$$
  zatem mamy do czynienia z rozkładem wykładniczym. Skumulowana funkcja hazardu wyraża się przez
  $$H(t) = ct,$$
  zaś funkcja przeżycia
  $$S(t) = \exp(-ct).$$
- *Funkcja liniowa* (afiniczna).
  Liniowość funkcji hazardu
  $$h(t) = \alpha + \beta t, \alpha, \beta > 0$$
  pociąga za sobą gęstość
  $$f(t) = (\alpha + \beta t) \exp\left(-\alpha t - \frac{\beta t^2}{2}\right).$$
  Skumulowana funkcja hazardu w tym przypadku to
  $$H(t) = \alpha t + \frac{\beta t^2}{2},$$

funkcja przeżycia natomiast

$$S(t) = \exp\left(-\alpha t - \frac{\beta t^2}{2}\right)$$

- *Funkcja wykładnicza.*
  Wykładnicza funkcja hazardu

  $$h(t) = \lambda \exp(\alpha t)$$

  determinuje funkcję gęstości będącą gęstością rozkładu Gompertza

  $$f(t) = \lambda \exp(\alpha t) \exp\left(\frac{\lambda}{\alpha}(1 - \exp(\alpha t))\right),$$

  skumulowaną funkcję hazardu

  $$H(t) = \frac{\lambda}{\alpha}(\exp(\alpha t) - 1)$$

  oraz funkcję przeżycia

  $$S(t) = \exp\left(\frac{\lambda}{\alpha}(1 - \exp(\alpha t))\right).$$

- *Funkcja potęgowa.*
  Dla funkcji hazardu danej przez

  $$h(t) = \lambda \nu t^{\nu-1}$$

  otrzymujemy rozkład Weibulla o gęstości

  $$f(t) = \lambda \nu t^{\nu-1} \exp(-\lambda t^{\nu}),$$

  skumulowanej funkcji hazardu

  $$H(t) = \lambda t^{\nu}$$

  i funkcji przeżycia

  $$\exp(-\lambda t^{\nu}).$$

4. Dane niepełne. W praktyce rzadko spotykamy się z sytuacją, gdy dane o wszystkich pacjentach biorących udział w badaniu są kompletne — z przyczyn niezależnych lub celowo pewne obserwacje mogą być niepełne. W pierwszym przypadku mówimy o danych cenzorowanych, w drugim — o danych obciętych. Dokładna ich charaktestyka oraz przykłady zostały opisane np. w [6].

4.1. *Dane cenzorowane.* Rozróżniamy trzy podstawowe typy cenzorowania: prawostronne, lewostronne oraz przedziałowe.

Cenzorowanie prawostronne, które stanowi najczęstszy typ cenzorowania, odnosi się do sytuacji, gdy obserwacja urywa się w pewnym momencie. Może to stać się w sposób *losowy*, gdy dzieje się przed końcem badania,

z powodów innych niż oczekiwane zdarzenie (np. przeprowadzka, wypadek samochodowy), będących poza kontrolą badającego lub *ustalony*, gdy przestajemy obserwować pacjenta z powodu zakończenia badania i jego śmierć w czasie późniejszym nie zostanie odnotowana.

Z *cenzorowaniem lewostronnym* spotykamy się, gdy interesujące nas zdarzenie zaszło zanim rozpoczęliśmy obserwację. Mamy tutaj do czynienia z procesem selekcji występującym losowo na poziomie indywidualnym. W tym przypadku interesujące nas zdarzenie miało miejsce we wcześniejszym, nieznanym nam czasie.

Przykładowo, modelując wiek rozpoczęcia regularnego palenia, możemy spotkać się z badanym, który nie pamięta, kiedy zaczął palić — wiemy więc tylko, że jego aktualny wiek jest większy niż ten, w którym miało to miejsce.

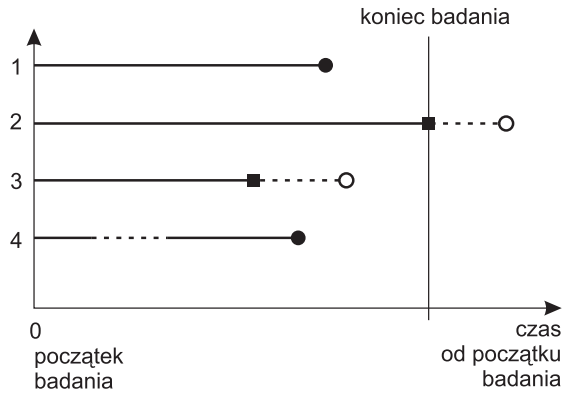
Można powiedzieć, że cenzorowanie lewostronne jest przeciwieństwem cenzorowania prawostronnego, gdzie wiemy, że zdarzenie nie miało jeszcze miejsca i że czas obserwowany jest mniejszy niż faktyczny czas przeżycia.

Możliwym rozwiązaniem (zaproponowanym w [8]) jest odwrócenie osi czasu i traktowanie danych jak cenzorowanych prawostronnie. Metoda ta działa jednak tylko, gdy występuje samo cenzorowanie lewostronne. W praktyce natomiast, jeśli zaobserwujemy lewostronne, dość prawdopodobnym jest, że prawostronne również wystąpi.

Dane cenzorowane przedziałowo są formą niekompletnych obserwacji, która może obejmować zarówno dane cenzorowane prawostronnie, lewostronnie, jak i obcięte (o których będzie mowa za chwilę). Powstają w sytuacji, gdy o czasie przeżycia wiemy tylko tyle, że leży pomiędzy dwoma wartościami.

Zdarza się to np. wówczas, gdy kontrola stanu badanych jest przeprowadzana co pewien ustalony odstęp czasu, np. w sytuacji, gdy pacjenci nie leżą w szpitalu i kontaktujemy się z nimi oraz sprawdzamy ich stan co trzy miesiące. Wtedy pacjenci, którzy zmarli w ciągu pierwszych trzech miesięcy stanowią obserwacje cenzorowane lewostronnie. Ci, którzy umrą między dwoma kolejnymi kontrolami, będą mieli czas przeżycia co najmniej taki, jak czas przedostatniej kontroli i mniejszy niż ostatniej. Czasy życia pacjentów żyjących w trakcie ostatniej kontroli będą natomiast cenzorowane prawostronnie, a więc co najmniej takie, jak czas ostatniego z nimi kontaktu.

W ustalonej chwili t jako obiekty będące w ryzyku zdarzenia traktujemy te, które w tym momencie znajdują się pod obserwacją. Dopóki obiekty te są reprezentatywne dla całej populacji, możemy na ich podstawie uzyskać nieobciążone estymatory prawdopodobieństwa przetrwania, czasu przeżycia i innych interesujących nas funkcji. Mechanizm cenzorowania powinien więc być niezależny od czasu przetrwania, mówiąc ściślej — rozkład czasów przetrwania obiektów cenzorowanych w pewnej chwili t nie powinien różnić się od rozkładu czasów przetrwania obiektów będących w tej chwili pod obser-



Rys. 1. Różne rodzaje cenzorowań: 1 — obserwacja niecenzorowana, 2 — cenzorowanie prawostronne ustalone, 3 — cenzorowanie prawostronne losowe, 4 — cenzorowanie przedziałowe.

wacją. Jeżeli warunek ten jest spełniony, mówimy, że cenzorowanie nie niesie informacji i z sytuacją taką spotykamy się przy cenzorowaniu ustalonym.

Natomiast w przypadku cenzorowania losowego może się zdarzyć (a nawet jest dość prawdopodobne), że czas przetrwania będzie w pewien sposób powiązany z mechanizmem cenzorowania. Przykładowo, pacjent w zaawansowanym stadium choroby, który na krótko przed śmiercią może zechcieć wycofać się z badania (i wówczas jego śmierć nie zostanie zaobserwowana i uwzględniona), zawyży czas przeżycia. Dlatego w badaniach, w których nie możemy wykluczyć pojawienia się cenzorowania losowego, w modelu uwzględniamy zmienne niezależne (objaśniające), które będą powiązane zarówno z cenzorowaniem, jak i z czasem przetrwania (np. stopień zaawansowania choroby na początku badania).

Łącznie dla danych cenzorowanych potrzebne nam są więc następujące informacje:

- czas przetrwania (bądź cenzorowania),
- czy czas przetrwania jest cenzorowany, czy też nie,
- na ogół jedna lub więcej zmiennych niezależnych mogących mieć wpływ na czas przeżycia (być może zależnych od czasu).

4.2. *Dane obcięte.* W przypadku, gdy niekompletność danych jest pochodną projektu badania, mówimy o *danych obciętych*. Mogą to być dane obcięte *lewostronnie*, gdy obiekty o czasie życia krótszym niż pewna ustalona wartość nie są w ogóle obserwowane (możemy o nich w ogóle nie wiedzieć, w przeciwnieństwie do lewostronnego cenzorowania — np. gdy nie obserwujemy dzieci do czasu, aż pójdą do szkoły), bądź *prawostronnie*, gdy cała obserwowana populacja doświadczyła interesującego nas zdarzenia przed rozpoczęciem badania.

Dane lewostronnie obcięte stanowią po cenzorowaniu prawostronnym najczęściej spotykany rodzaj danych niepełnych.

Możemy spotkać się z nimi np. w sytuacji, gdy zaczynamy obserwować pacjentów dopiero gdy ukończą oni pewien wstępny program leczenia, ale czas przeżycia mierzymy od chwili przystąpienia do tego programu — mamy tu więc do czynienia z procesem selekcji, który kwalifikuje pacjentów do badania. Ich minimalnym czasem przetrwania będzie w związku z tym długość leczenia wstępnego, co sprawia, że czasy przeżycia dla wszystkich pacjentów biorących udział w badaniu przekraczają pewną ustaloną wartość. Pozostałe będą wykluczone.

W tym przypadku mówimy też o opóźnionym wejściu do badania (ang. *delayed entry*): fakt przetrwania obiektu w badaniu z danymi lewostronnie obciętymi nie jest poddawany analizie do czasu, gdy tempo hazardu (estymowane przez stosunek liczby zdarzeń do liczby będących w ryzyku) nie przekroczy pewnej określonej liczby — wejście do zbioru ryzyka jest zatem odłożone w czasie (i jeśli obiekt zostanie do niego włączony, pozostaje w nim aż do czasu wystąpienia oczekiwanego zdarzenia bądź np. prawostronnego cenzorowania).

Z punktu widzenia samego modelu i użytej w nim struktury danych każdy obiekt w badaniu można opisać przez wartość początkową czasu (niekoniecznie 0), czas końca badania oraz zmienną wskazującą, czy obserwacja jest prawostronnie cenzorowana. Możemy w ten sposób otrzymać wszystkie możliwe kombinacje lewostronnego obcięcia i prawostronnego cenzorowania.

Dane prawostronnie obcięte pojawiają się również w kontekście selekcji — gdy mamy dane jedynie od obiektów, które doświadczyły interesującego nas zdarzenia. Przykładem takich danych mogą być wszelkie rejestry, zawierające informacje o udokumentowanych przypadkach choroby.

Dlatego też jakakolwiek analiza wykorzystująca potwierdzone przypadki zachorowań (gdzie taki przypadek jest badanym przez nas wydarzeniem) będzie pociągała za sobą prawostronne obcięcie.

5. Modele hazardu. Modele hazardu — w tym model Coxa — zostały dość szczegółowo opisane np. w [6].

5.1. *Wprowadzenie.* Wyróżniającą na tle innych zmiennych zależnych cechą czasu przeżycia jest fakt nieodłącznego starzenia się badanego obiektu w czasie. Esencję tego procesu najlepiej oddaje funkcja hazardu i to ją będziemy chcieli umieścić w tworzonym modelu regresji.

Modelując czas przeżycia możemy mieć dwa cele — opisanie jego podstawowego rozkładu oraz scharakteryzowanie, jak ów rozkład zmienia się jako funkcja zmiennych niezależnych. Oba powinny być zrealizowane, jeżeli np. badamy czas działania dysku twardego jako funkcję temperatury i wil-

gotności powietrza — szukany model ma za zadanie przewidzieć czas życia urządzenia w określonych warunkach użytkowania.

Gdy natomiast chcemy ocenić, czy połączenie dwóch terapii poprawia szanse na przeżycie pacjentów w stosunku do pojedynczej terapii, bardziej niż na dokładnym opisie rozkładu czasu przeżycia zależy nam na sprawdzeniu skuteczności nowego sposobu leczenia na tle poprzedniego. Modele stosowane do opisywania czasu życia w sensie porównawczym noszą nazwę *modeli semiparametrycznych*.

5.2. *Zmienne niezależne.* Zmienne niezależne (objaśniające) są to zmienne, na podstawie których wyliczać będziemy wartości zmiennych zależnych (objaśnianych).

W zależności od kierunku działania tych zmiennych można wyróżnić wśród nich *czynniki ryzyka*, czyli zmienne mogące wpływać na pojawienie się dolegliwości (np. liczba wypalanych dziennie papierosów), *symptomy* — efekty dolegliwości (np. podwyższony poziom cukru we krwi u chorych na cukrzycę), oraz zmienne będące czynnikami ryzyka i symptomami jednocześnie (np. wysoki poziom cholesterolu mogący powodować choroby serca może być również symptomem takiego schorzenia).

Zmienne te mogą być zarówno dyskretne (w tym np. wskaźnikowe, przyporządkowujące każdemu obiektowi numer grupy, do której należy, jak też inne — liczba wcześniejszych ataków choroby czy hospitalizacji), jak i ciągłe (np. wiek czy poziom choresterolu).

5.3. *Modele parametryczne i semiparametryczne.* Chcąc opisać rozkład czasu życia możemy wykorzystać jedną z dwóch funkcji — gęstość tego rozkładu lub jego funkcję hazardu. Zaletą drugiej z nich jest oddawanie wprost procesu starzenia się badanego obiektu, dlatego też w konstruowanym modelu będziemy starali się badać zależność między przeżyciem wyrażającym się właśnie przez funkcję hazardu a pewnymi zmiennymi niezależnymi.

Do najpopularniejszych modeli należą log-liniowe modele hazardu. Przykładem takiego modelu jest (oparty na rozkładzie wykładniczym)

$$\ln h_i(t) = \alpha + \beta_1 x_{i1} + \ldots + \beta_k x_{ik},$$

gdzie i — indeks obiektu. Jest to model liniowy dla log-hazardu (bądź multiplikatywny dla samego hazardu) oraz *parametryczny*, ponieważ mając wyznaczone parametry regresji α, β_j, jednoznacznie charakteryzujemy nimi funkcję hazardu. Stała regresji α odpowiada tzw. hazardowi bazowemu, tzn. sytuacji, gdy wszystkie zmienne x_{ij} są równe 0.

Model w pełni parametryczny można jednak zastąpić modelem, w którym hazard bazowy $\alpha(t) = \ln h_0(t)$ pozostaje nieokreślony i może mieć jakąkolwiek formę, zaś zmienne objaśniające wchodzą do niego przez liniowy predyktor $\eta_i = \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$, nie zawierający stałego czynnika (wchło-

niętego przez hazard bazowy). Wówczas mamy do czynienia z *modelem semiparametrycznym*:

$$\ln h_i(t) = \alpha(t) + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}.$$

5.4. *Modele proporcjonalne.* Model proporcjonalny jest to model, w którym funkcja hazardu, określona jako funkcja czasu i zmiennych objaśniających, przyjmuje postać

$$h(t, x, \beta) = h_0(t) r(x, \beta). \tag{5.1}$$

Jak widać, jest ona iloczynem dwóch funkcji: wspomnianego wcześniej hazardu bazowego $h_0(t)$ określającego, jak hazard zmienia się jako funkcja czasu oraz $r(x, \beta)$ mówiącej, jak się zmienia jako funkcja zmiennych niezależnych. Funkcje te muszą być tak dobrane, aby $h(t, x, \beta) > 0$.

Stosunek funkcji hazardu dla dwóch obiektów o zmiennych objaśniających x_1 i x_2 w takim modelu wynosi

$$HR(t, x_1, x_2) = \frac{h(t, x_1, \beta)}{h(t, x_2, \beta)} = \frac{h_0(t) r(x_1, \beta)}{h_0(t) r(x_2, \beta)} = \frac{r(x_1, \beta)}{r(x_2, \beta)}, \tag{5.2}$$

zatem współczynnik hazardu HR zależy jedynie od funkcji zmiennych niezależnych. Dlatego — o ile będzie on łatwy w interpretacji — forma hazardu bazowego będzie miała dla nas niewielkie znaczenie.

5.5. *Model Coxa*

5.5.1. Postać modelu. David Cox (1972) wprowadził model, w którym funkcja zmiennych niezależnych wyraża się wzorem

$$r(x, \beta) = e^{x\beta}.$$

Funkcja hazardu w tym modelu przybiera więc postać

$$h(t, x, \beta) = h_0(t) e^{x\beta}, \tag{5.3}$$

zaś współczynnik hazardu określony jest jako

$$HR(t, x_1, x_2) = e^{\beta(x_1 - x_2)}.$$

W literaturze model ten spotyka się pod nazwami model Coxa (*Cox model*), model proporcjonalnych hazardów Coxa (*Cox proportional hazards model*) lub model hazardów proporcjonalnych (*proportional hazards model*).

Jest to najczęściej używany typ modelu proporcjonalnego semiparametrycznego.

5.5.2. Współczynnik hazardu. $HR(t, x_1, x_2)$ w modelu Coxa interpretuje się jako współczynnik ryzyka względnego. Przykładowo, dla zmiennej niezależnej dychotomicznej, jak np. płeć, z wartościami $x_1 = 1$ dla mężczyzn i $x_2 = 0$ dla kobiet, współczynnik hazardu wynosi

$$HR(t, x_1, x_2) = e^{\beta}.$$

Dla $\beta = \ln(2)$ otrzymujemy stąd, że tempo umieralności mężczyzn w tym przypadku jest dwukrotnie większe od tempa umieralności kobiet.

5.5.3. Funkcja przeżycia. Ogólnie funkcja przeżycia wyraża się wzorem

$$S(t, x, \beta) = e^{-H(t,x,\beta)},$$

gdzie $H(t, x, \beta)$ jest skumulowaną funkcją hazardu w chwili t dla obiektu o zmiennej niezależnej x. W modelu proporcjonalnym:

$$H(t, x, \beta) = \int_0^t h(u, x, \beta)du = r(x, \beta) \int_0^t h_0(u)du = r(x, \beta)H_0(t),$$

zatem łącznie

$$S(t, x, \beta) = e^{-r(x,\beta)H_0(t)}. \tag{5.4}$$

Możemy to inaczej zapisać jako

$$S(t, x, \beta) = [e^{-H_0(t)}]^{r(x,\beta)} = [S_0(t)]^{r(x,\beta)},$$

gdzie $S_0(t) = e^{-H_0(t)}$ jest *bazową funkcją przeżycia*.

W samym modelu Coxa funkcja przeżycia określona jest więc wzorem

$$S(t, x, \beta) = [S_0(t)]^{\exp(x\beta)}. \tag{5.5}$$

Przyjmuje ona oczywiście wartości między 0 a 1.

P r z y k ł a d 5.1 [6]. Niech zmienną niezależną będzie wiek badanego pacjenta a. Oznaczmy przez $x = a - \bar{a}$. Załóżmy, że ryzyko zapadnięcia na pewną chorobę jest związane z wiekiem ($\beta > 0$). Wówczas dla pacjenta w wieku dokładnie $\bar{a}$:

$$S(t, x, \beta) = S_0(t).$$

Dla pacjenta w wieku powyżej średniej ($a > \bar{a}$) mamy

$$x > 0,$$

co daje

$$e^{x\beta} > 1$$

i w konsekwencji

$$S(t, x, \beta) < S_0(t).$$

Oznacza to, że (zgodnie z intuicją) pacjent w wieku powyżej średniej będzie miał mniejsze szanse na przeżycie.

Dla pacjentów młodszych natomiast ($a < \bar{a}$) mamy

$$x < 0,$$

a więc

$$e^{x\beta} < 1$$

i w związku z tym

$$S(t, x, \beta) > S_0(t),$$

co wskazuje na większą szansę przeżycia.

5.5.4. Zalety modelu. Model Coxa posiada wiele niewątpliwych zalet. W odróżnieniu od klasycznych modeli regresji można go stosować, gdy zmienna zależna nie ma rozkładu normalnego oraz gdy mamy do czynienia z danymi niepełnymi (o czym będzie mowa później).

Ponadto, ponieważ zawsze

$$e^{x\beta} > 0,$$

nie ma potrzeby nakładania dodatkowych założeń na wartości wyrażenia $x\beta$.

Dodatkowo, założenie proporcjonalności implikuje, że znajomość jedynie współczynników β_i, bez wiedzy o $h_0(t)$, pozwala określić wrażliwość funkcji hazardu na zmianę konkretnej cechy.

6. Estymacja parametrów modelu. Metody estymacji parametrów modelu Coxa — w różnych przypadkach i przy różnych założeniach — zostały opisane m.in. w [3] i [6].

6.1. *Przypadek z hazardem bazowym przedziałami stałym.* Na początku rozważymy przypadek uproszczony, w którym funkcja hazardu bazowego jest przedziałami stała.

6.1.1. Oznaczenia. Zakładamy, że

$$h_0(t) = h_k,$$

gdzie:

$t \in I_k = (t_{k-1}, t_k]$,
$k = 1, \ldots, M$.

Oznaczmy dalej:

$\mathcal{S}_k$ – zbiór osób będących pod obserwacją kiedykolwiek w przedziale I_k,
I_{kl} – podprzedział I_k, w którym osoba l była pod obserwacją,
d_{kl} – długość przedziału I_{kl},
$d_k = t_k - t_{k-1}$ – długość przedziału I_k.

Jeżeli nie znamy dokładnego czasu wycofania się danej osoby z badania, możemy aproksymować d_{kl} przez $\frac{1}{2}(t_k - t_{k-1})$ bądź $\frac{1}{3}(t_k - t_{k-1})$, jeśli rozpoczęcie i zakończenie obserwacji danej osoby miało miejsce w tym samym przedziale I_k.

Całkując funkcję hazardu otrzymujemy

$$H_k(x_l) = h_k \int\limits_{I_{kl}} r(x_l; \beta)dt = h_k d_{kl} r(x_l; \beta)$$

(funkcja r nie zależy od czasu t).

Niech ponadto

$$\delta_{kl} = \begin{cases} 1 & \text{jeżeli } l\text{-ty badany zmarł w przedziale } I_k \\ 0 & \text{w przeciwnym przypadku.} \end{cases}$$

6.1.2. Funkcja wiarygodności. Wkład obserwacji w przedziale I_k do funkcji wiarygodności wyraża się wzorem

$$L_k(h_k;\beta) = \prod_{l\in\mathcal{S}_k} [h_k r(x_l;\beta)]^{\delta_{kl}} \exp[-h_k d_{kl} r(x_l;\beta)],$$

zatem pełna funkcja wiarygodności ma postać:

$$L(h_1,\ldots,h_M;\beta) = \prod_{k=1}^{M} L_k(h_k;\beta)). \tag{6.1}$$

6.1.3. Estymator największej wiarygodności. Estymator największej wiarygodności dla h_k jest równy

$$\hat{h_k}(\beta) = \frac{\sum_{l\in\mathcal{S}_k} \delta_{kl}}{\sum_{l\in\mathcal{S}_k} d_{kl} r(x_l;\beta)}. \tag{6.2}$$

Licznik tego wyrażenia jest równy łącznej liczbie zgonów w przedziale I_k. Jeżeli zatem w danym przedziale czasowym nikt nie umiera, wówczas $\hat{h_k} = 0$, niezależnie od wartości β.

Jeżeli w modelu nie ma zmiennych objaśniających (np. $r(x_l;\beta) \equiv 1$), wtedy estymator funkcji hazardu wyraża się po prostu wzorem

$$\widehat{h_k} = \frac{\text{liczba zgonów}}{\text{liczba wystawionych na ryzyko}}.$$

Podstawiając estymator h_k do wzoru na $L_k(h_k;\beta)$ otrzymujemy maksymalną wartość tego czynnika przy danym β:

$$\begin{aligned} \hat{L_k}(\beta) = \prod_{l\in\mathcal{S}_k} & \left[\frac{\sum_{l'\in\mathcal{S}_k} \delta_{kl'}}{\sum_{l'\in\mathcal{S}_k} d_{kl'} r(x'_l;\beta)} r(x_l;\beta)\right]^{\delta_{kl}} \cdot \\ & \cdot \exp\left[-\frac{\sum_{l'\in\mathcal{S}_k} \delta_{kl'}}{\sum_{l'\in\mathcal{S}_k} d_{kl'} r(x'_l;\beta)} d_{kl'} r(x'_l;\beta)\right] = \\ = & \left(\frac{\sum_{l'\in\mathcal{S}_k} \delta_{kl'}}{\sum_{l'\in\mathcal{S}_k} d_{kl'} r(x'_l;\beta)}\right)^{\sum_{l\in\mathcal{S}_k} \delta_{kl}} \cdot \\ & \cdot \exp\left(-\sum_{l\in\mathcal{S}_k} \delta_{kl}\right) \prod_{l\in\mathcal{S}_k} r(x_l;\beta)^{\delta_{kl}} = \end{aligned} \tag{6.3}$$

$$= \left[e^{-1} \frac{\sum_{l' \in \mathcal{S}_k} \delta_{kl'}}{\sum_{l' \in \mathcal{S}_k} d_{kl'} r(x'_l; \beta)} \right]^{\sum_{l \in \mathcal{S}_k} \delta_{kl}} \prod_{l \in \mathcal{S}_k^*} r(x_l; \beta),$$

gdzie $\mathcal{S}_k^*$ jest zbiorem osób zmarłych w przedziale I_k.

Estymator największej wiarygodności dla β musi zatem maksymalizować funkcję

$$L(\hat{\beta}) = \prod_{k=1}^{M} \hat{L_k}(\hat{\beta}).$$

Na ogół wartości $\hat{\beta}$ oblicza się korzystając z metod numerycznych.

6.1.4. Dobór przedziałów I_k. Warto zwrócić uwagę, że im krótsze przedziały, tym bliższa aproksymacja $h_0(t)$, ale jednocześnie liczba obserwacji w przedziale staje się mniejsza, co skutkuje mniejszą dokładnością estymatora.

Jeżeli założymy, że czasy śmierci są parami różne, otrzymujemy w granicy wzór:

$$\hat{h_j} = \hat{h_j}(\beta) = \frac{1}{d_j \sum_{l \in \mathcal{R}_j} r(x_l; \beta)},$$

gdzie $d_j = t'_j - t'_{j-1}$ jest czasem, który upłynął między $(j-1)$. a j. zgonem, zaś $\mathcal{R}_j$ jest zbiorem osób żyjących tuż przed czasem t'_j.

Estymator największej wiarygodności dla β maksymalizuje wartość wyrażenia

$$\prod_{j=1}^{n} d_j e^n \hat{L}(\beta) = \prod_{j=1}^{n} \frac{r(x_{i(j)}; \beta)}{\sum_{l \in \mathcal{R}_j} r(x_l; \beta)},$$

gdzie n jest całkowitą liczbą zgonów, zaś $i(j)$ oznacza indeks osoby, która zmarła jako $j-$ta.

W powyższych wzorach można jeszcze zastąpić $\sum_{l \in \mathcal{R}_j}$ wygodniejszą formą. Niech τ_l oznacza czas zakończenia obserwacji (spowodowanej śmiercią bądź wycofaniem się z badania) osoby l-tej. Zdefiniujmy

$$\theta_{jl} = \begin{cases} 0 & \text{jeżeli } \tau_l < t'_j \\ 1 & \text{jeżeli } \tau_l \geqslant t'_j. \end{cases}$$

Wtedy

$$\sum_{l \in \mathcal{R}_j} r(x_l; \beta) = \sum_{l=1}^{N} \theta_{jl} r(x_l; \beta).$$

6.1.5. Estymacja w modelu Coxa. W modelu Coxa, uwzględniając jeszcze ostatnie oznaczenia, otrzymujemy ostatecznie estymatory:

$$h_j = h_j(\beta) = \frac{1}{h_j \sum_{l=1}^{N} \theta_{jl} \exp(\beta' x_l)} \tag{6.4}$$

oraz

$$\prod_{j=1}^{n} d_j e^n \hat{L}(\beta) = \prod_{j=1}^{n} \frac{\exp(\beta' x_l)}{\sum_{l=1}^{N} \theta_{jl} \exp(\beta' x_l)}. \tag{6.5}$$

6.2. *Przypadek ogólny — funkcja wiarygodności.* W przypadku ogólnym ([6]) dla każdego z n obserwowanych obiektów dysponujemy początkowo trójką

$$(t_i, x_i, c_i),$$

gdzie:

t_i – długość obserwacji,

x_i – zmienna niezależna (jedna, jej wartość jest zdeterminowana na początku obserwacji i pozostaje niezmienna przez cały czas trwania badania),

c_i – informacja, czy obserwacja jest cenzorowana, czy też nie ($c_i = 0$, gdy powodem zakończenia obserwacji nie było zdarzenie, króre nas interesuje, 1 w przeciwnym przypadku).

Funkcja wiarygodności, za pomocą której będziemy estymować parametry modelu, ma przy powyższych oznaczeniach postać:

$$l(\beta) = \prod_{i=1}^{n} \Big\{ [f(t_i, x_i, \beta)]^{c_i} \times [S(t_i, x_i, \beta)]^{1-c_i} \Big\}.$$

Korzystając z zależności

$$f(t, x, \beta) = h(t, x, \beta) S(t, x, \beta) \tag{6.6}$$

otrzymujemy, że

$$l(\beta) = \prod_{i=1}^{n} \Big\{ [h(t_i, x_i, \beta) S(t_i, x_i, \beta)]^{c_i} \times [S(t_i, x_i, \beta)]^{1-c_i} \Big\},$$

co po uproszczeniu daje

$$l(\beta) = \prod_{i=1}^{n} \{ [h(t_i, x_i, \beta)]^{c_i} \times [S(t_i, x_i, \beta)] \}.$$

Estymatorem dla β będzie oczywiście wartość minimalizująca tę funkcję. Biorąc z niej logarytm i podstawiając konkretną postać funkcji hazardu otrzymujemy

$$L(\beta) = \sum_{i=1}^{n} \left\{ c_i ln[h_0(t_i)] + c_i x_i \beta + e^{x_i \beta} \ln [S_0(t_i)] \right\}. \tag{6.7}$$

6.3. *Funkcja częściowej wiarygodności.* Użycie funkcji częściowej wiarygodności, która będzie zależałą jedynie od interesującego nas parametru,

zostało zaproponowane przez Coxa ([6]). Przypuszczał on, że parametry wyestymowane w ten sposób będą miały takie same własności „rozkładowe" jak te otrzymane pełną metodą największej wiarygodności. Formalnie udowodniono to później.

Funkcja ta ma postać

$$l_p(\beta) = \prod_{i=1}^{n} \left[\frac{e^{x_i\beta}}{\sum_{j\in R(t_i)} e^{x_j\beta}} \right]^{c_i}, \tag{6.8}$$

gdzie $R(t_i)$ (od ang. *risk set*) zawiera wszystkie obiekty o czasie życia (lub czasie cenzorowania) większym bądź równym t_i, czyli wystawione w chwili t_i na ryzyko.

Wyrażenie to często modyfikuje się opuszczając składniki z $c_i = 0$, co prowadzi do

$$l_p(\beta) = \prod_{i=1}^{m} \frac{e^{x_{(i)}\beta}}{\sum_{j\in R(t_{(i)})} e^{x_j\beta}}, \tag{6.9}$$

gdzie mnożenie odbywa się po m różnych, uporządkowanych czasach przeżycia, zaś $x_{(i)}$ oznacza wartość zmiennej niezależnej dla obiektu z czasem $t_{(i)}$ (przy założeniu, że nie mamy danych zaokrąglanych, o czym będzie jeszcze mowa później).

Współczynnik

$$\frac{e^{x_i\beta}}{\sum_{j\in R(t_i)} e^{x_j\beta}}$$

ma interpretację intuicyjną. Hazard dla obiektu i (którego czas przetrwania jest równy t_i) jest proporcjonalny do $e^{x_i\beta}$. Powyższy współczynnik wyraża hazard dla obiektu i w stosunku do sumy hazardów wszystkich obiektów będących w ryzyku w momencie, w którym i-ty doświadcza zdarzenia. Chcemy, aby wartości wyestymowanych parametrów były takie, by hazard był wysoki dla obiektu w chwili, gdy interesujące nas zdarzenie faktycznie mu się przytrafia.

Dalej postępujemy jak przy pełnej funkcji wiarygodności, tzn. bierzemy logarytm z powyższego wyrażenia:

$$L_p(\beta) = \sum_{i=1}^{m} \left\{ x_{(i)}\beta - \ln \left[\sum_{j\in R(t_{(i)})} e^{x_j\beta} \right] \right\}$$

i różniczkujemy:

$$\frac{\partial L_p(\beta)}{\partial \beta} = \sum_{i=1}^{m} \left\{ x_{(i)} - \frac{\sum_{j\in R(t_{(i)})} x_j e^{x_j\beta}}{\sum_{j\in R(t_{(i)})} e^{x_j\beta}} \right\} \tag{6.10}$$

$$= \sum_{i=1}^{m} \left\{ x_{(i)} - \sum_{j \in R(t_{(i)})} w_{ij}(\beta) x_j \right\}$$
$$= \sum_{i=1}^{m} x_{(i)} - \bar{x}_{w_i},$$

gdzie

$$w_{ij}(\beta) = \frac{e^{x_j \beta}}{\sum_{l \in R(t_{(i)})} e^{x_l \beta}},$$

zaś

$$\bar{x}_{w_i} = \sum_{j \in R(t_{(i)})} w_{ij}(\beta) x_j.$$

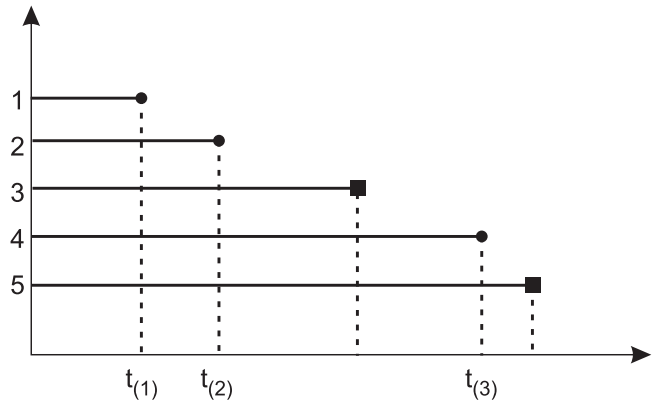
Otrzymane w ten sposób wyrażenie przyrównujemy do 0 i rozwiązujemy równanie.

Bez opuszczenia składników z $c_i = 0$ pochodna logarytmu funkcji częściowej wiarygodności ma postać

$$\frac{\partial L_p(\beta)}{\partial \beta} = \sum_{i=1}^{n} c_i \left\{ x_i - \frac{\sum_{j \in R(t_{(i)})} x_j e^{x_j \beta}}{\sum_{j \in R(t_{(i)})} e^{x_j \beta}} \right\}. \tag{6.11}$$

Otrzymany tą drogą estymator — rozwiązanie tych równań — oznaczamy przez $\hat{\beta}$.

P r z y k ł a d 6.1. Przypuśćmy, że w badaniu brało udział pięciu pacjentów, których czasy przeżycia bądź cenzorowania przedstawia wykres na rysunku 2.



Rys. 2. Czasy przeżycia pacjentów (kwadrat oznacza obserwację cenzorowaną).

Wartości zmiennych objaśniających określone są w tabelce 6.1.

	1	2	3	4	5
x	10	10	20	20	10

Tabela 6.1. Wartości zmiennych objaśniających.

Z wykresu na rysunku 2. wynika, że:

- pacjenci numer 1, 2 i 4 zmarli odpowiednio w chwilach $t_{(1)}$, $t_{(2)}$ i $t_{(3)}$,
- pacjenci numer 3 i 5 opuścili badanie z przyczyn losowych (między $t_{(2)}$ a $t_{(3)}$ i po czasie $t_{(3)}$ odpowiednio), nie jest dla nas istotne, kiedy dokładnie miało to miejsce.

Z związku z powyższym mamy 3 zbiory ryzyka:
$R(t_{(1)}) = \{1, 2, 3, 4, 5\}$
$R(t_{(2)}) = \{2, 3, 4, 5\}$
$R(t_{(3)}) = \{4, 5\}$

Łącznie dane potrzebne do obliczenia funkcji częściowej wiarygodności przedstawia tabelka 6.2.

i	x_i	$e^{x_i\beta}$	(i)	$R(t_{(i)})$	$\sum_{j\in R(t_{(i)})} e^{x_j\beta}$
1	10	$e^{10\beta}$	1	$\{1, 2, 3, 4, 5\}$	$3e^{10\beta} + 2e^{20\beta}$
2	10	$e^{10\beta}$	2	$\{2, 3, 4, 5\}$	$2e^{10\beta} + 2e^{20\beta}$
3	20	$e^{20\beta}$	–	–	–
4	20	$e^{20\beta}$	3	$\{4, 5\}$	$e^{10\beta} + e^{20\beta}$
5	10	$e^{10\beta}$	–	–	–

Tabela 6.2. Wartości zmiennych objaśniających i odpowiadające im składniki funkcji częściowej wiarygodności.

Podstawiamy je do wzoru:

$$l_p(\beta) = \prod_{i=1}^{3} \frac{e^{x_{(i)}\beta}}{\sum_{j\in R(t_{(i)})} e^{x_j\beta}}$$

i otrzymujemy:

$$l_p(\beta) = \frac{e^{10\beta}}{3e^{10\beta} + 2e^{20\beta}} \times \frac{e^{10\beta}}{2e^{10\beta} + 2e^{20\beta}} \times \frac{e^{20\beta}}{e^{10\beta} + e^{20\beta}}$$

lub, po zlogarytmowaniu i zróżniczkowaniu:

$$\frac{\partial L_p(\beta)}{\partial\beta} = 40 - \left(\frac{30e^{10\beta} + 40e^{20\beta}}{3e^{10\beta} + 2e^{20\beta}} + \frac{20e^{10\beta} + 40e^{20\beta}}{2e^{10\beta} + 2e^{20\beta}} + \frac{10e^{10\beta} + 20e^{20\beta}}{e^{10\beta} + e^{20\beta}} \right),$$

co przyrównane do 0 pozwala wyestymować β. Otrzymujemy, że

$$\hat{\beta} = -0.0564.$$

6.4. *Estymator wariancji.* Estymator wariancji dla estymatora współczynnika β otrzymujemy postępując podobnie, jak w większości zastosowań metody największej wiarygodności.

Liczymy drugą pochodną z funkcji częściowej wiarygodności:

$$\frac{\partial^2 L_p(\beta)}{\partial \beta^2} = \tag{6.12}$$

$$-\sum_{i=1}^{m} \left\{ \frac{\left[\sum_{j\in R(t_{(i)})} e^{x_j\beta}\right]\left[\sum_{j\in R(t_{(i)})} x_j^2 e^{x_j\beta}\right] - \left[\sum_{j\in R(t_{(i)})} x_j e^{x_j\beta}\right]^2}{\left[\sum_{j\in R(t_{(i)})} e^{x_j\beta}\right]^2} \right\}.$$

Wyrażenie to można uprościć stosując (jak wcześniej) oznaczenie $w_{ij}(\beta)$ (6.10):

$$\frac{\partial^2 L_p(\beta)}{\partial \beta^2} = -\sum_{i=1}^{m} \sum_{j\in R(t_{(i)})} w_{ij}(x_j - \bar{x}_{w_i})^2.$$

Wyrażenie przeciwne do powyższego (minus drugą pochodną) nazywamy *informacją zaobserwowaną* i oznaczamy przez $\mathbb{I}(\beta)$:

$$\mathbb{I}(\beta) = -\frac{\partial^2 L_p(\beta)}{\partial \beta^2}. \tag{6.13}$$

Estymatorem wariancji estymowanego współczynnika jest odwrotność informacji zaobserwowanej wyliczona w $\widehat{\beta}$:

$$\widehat{Var}(\widehat{\beta}) = \mathbb{I}(\widehat{\beta})^{-1}, \tag{6.14}$$

natomiast estymatorem błędu standardowego (ozn. $\widehat{SE}(\widehat{\beta})$) jest pierwiastek kwadratowy z estymatora wariancji:

$$\widehat{SE}(\widehat{\beta}) = \sqrt{\widehat{Var}(\widehat{\beta})}. \tag{6.15}$$

6.5. *Badanie istotności współczynnika.* Po wyestymowaniu współczynnika można ocenić jego istotność, posługując się jednym z poniższych testów.

6.5.1. Test współczynnika częściowej wiarygodności. Wartość oznaczaną przez G obliczamy mnożąc przez 2 różnicę logarytmu funkcji częściowej wiarygodności dla modelu zawierającego daną zmienną i nie zawierającego jej:

$$G = 2\left\{L_p(\widehat{\beta}) - L_p(0)\right\},$$

gdzie

$$L_p(0) = -\sum_{i=1}^{m} \ln(n_i),$$

przy czym n_i oznacza liczbę obiektów w zbiorze ryzyka w chwili $t_{(i)}$.

Przy hipotezie zerowej, że badany współczynnik jest równy 0, jest to statystyka o rozkładzie χ^2 o 1 stopniu swobody.

6.5.2. Test Walda. Obliczamy stosunek wyestymowanego współczynnika do jego błędu standardowego:

$$z = \frac{\widehat{\beta}}{\widehat{SE}(\widehat{\beta})}.$$

Statystyka ta będzie miała standardowy rozkład normalny.

6.5.3. Score test. Statystyka testowa jest stosunkiem pochodnej funkcji częściowej wiarygodności do pierwiastka informacji zaobserwowanej, obliczonych w punkcie $\beta = 0$:

$$z^* = \left.\frac{\frac{\partial L_p}{\partial \beta}}{\sqrt{\mathbb{I}(\beta)}}\right|_{\beta=0}.$$

Przy hipotezie, że badany współczynnik jest równy 0, statystyka ta ma standardowy rozkład normalny.

Wartości wszystkich trzech testów (G,z i z^*) powinny być zbliżone i prowadzić do tego samego wniosku. W sytuacji, gdy otrzymane wyniki są różne, preferuje się test współczynnika częściowej wiarygodności.

6.6. *Estymacja przy większej liczbie zmiennych niezależnych.* Opisywany do tej pory model zawierał tylko jedną zmienną niezależną. Można go jednak odpowiednio przeformułować, aby uwzględniał jednoczesne działanie wielu takich zmiennych.

Rozważmy zatem p zmiennych, których wartości są mierzone dla każdego obiektu na początku badania i nie zmieniają się z upływem czasu. i-temu obiektowi będzie więc odpowiadał p-wymiarowy wektor zmiennych niezależnych

$$\boldsymbol{x}'_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$$

i będziemy poszukiwali dla niego wektora współczynników

$$\beta' = (\beta_1, \beta_2, \ldots, \beta_p).$$

Wszystkie obliczenia będą przebiegać analogicznie do przypadku z jedną zmienną, z tą różnicą, że w miejsce x podstawimy powyższy wektor i otrzymamy tym samym układ p równań. Pochodna po k-tej zmiennej będzie miała postać:

$$\frac{\partial L_p(\beta)}{\partial \beta_k} = \sum_{i=1}^{m} \left\{ x_{ik} - \frac{\sum_{j \in R(t_{(i)})} x_{jk} e^{\boldsymbol{x}'_j \beta}}{\sum_{j \in R(t_{(i)})} e^{\boldsymbol{x}'_j \beta}} \right\} = \sum_{i=1}^{m} \{ x_{(ik)} - \bar{x}_{w_i k} \}, \tag{6.16}$$

gdzie

$$\bar{x}_{w_i k} = \sum_{j \in R(t_{(i)})} w_{ij}(\beta) x_{jk}$$

oraz

$$w_{ij}(\beta) = \frac{e^{\boldsymbol{x}'_j \beta}}{\sum_{l \in R(t_{(i)})} e^{\boldsymbol{x}'_l \beta}}.$$

Przez $x_{(ik)}$ natomiast oznaczamy wartość zmiennej x_k dla obiektu z (uporządkowanym) czasem przeżycia $t_{(i)}$.

Estymatorem otrzymanym metodą największej wiarygodności będzie wektor $\widehat{\beta}' = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)$.

Podobnie, elementy *macierzy informacyjnej* (o wymiarach $p \times p$) otrzymujemy przez wyliczenie pochodnych cząstkowych drugiego rzędu:

$$\mathbb{I}(\beta) = -\frac{\partial^2 L(\beta)}{\partial \beta^2}, \tag{6.17}$$

na diagonali mamy więc

$$\frac{\partial^2 L_p(\beta)}{\partial \beta_k^2} = -\sum_{i=1}^{m} \sum_{j \in R(t_{(i)})} w_{ij}(x_{jk} - \bar{x}_{w_i k}),$$

zaś poza nią:

$$\frac{\partial^2 L_p(\beta)}{\partial \beta_k \partial \beta_l} = -\sum_{i=1}^{m} \sum_{j \in R(t_{(i)})} w_{ij}(x_{jk} - \bar{x}_{w_i k})(x_{jl} - \bar{x}_{w_i l}).$$

Estymator macierzy kowariancji estymatora największej częściowej wiarygodności otrzymujemy analogicznie — obliczając odwrotność macierzy informacyjnej w punkcie estymatora, tj.

$$\widehat{Var}(\widehat{\beta}) = \mathbb{I}(\widehat{\beta})^{-1}.$$

6.7. *Estymacja przy zaokrąglanych danych.* Skonstruowana wcześniej funkcja częściowej wiarygodności jest oparta na założeniu, że nie mamy *danych zaokrąglanych*, tzn. dotychczas przyjmowaliśmy, że każdy czas przeżycia jest unikalny i w danej chwili t co najwyżej jeden obiekt doświadcza interesującego nas zdarzenia.

Konstruując nową funkcję zakładamy, że zaokrąglenia dla konkretnego czasu przeżycia pojawiają się przez brak precyzji w pomiarach czasu przeżycia. W związku z tym, gdy mamy d wartości zaokrąglonych do jednej, w rzeczywistości mogły one zostać zaobserwowane w którejkolwiek z $d!$ możliwych kolejności.

Dokładna postać funkcji częściowej wiarygodności jest uzyskiwana przez modyfikację mianownika poprzedniej tak, by zawierał wszystkie możliwe uporządkowania. Wówczas jednak otrzymujemy wyrażenia niewygodne do

dalszych obliczeń, dlatego posługujemy się przybliżeniami. Aproksymacja (zarówno wprowadzona przez Breslowa (1974), jak i Efrona (1977)) dostarcza wyrażeń prostszych niż dokładna funkcja, ale wciąż uwzględniających wpływ zaokrąglanych danych.

Dla wygody zapisu rozważamy model z tylko jedną zmienną niezależną.

6.7.1. Aproksymacja Breslowa. Funkcję częściowej wiarygodności można aproksymować przez

$$l_{p1}(\beta) = \prod_{i=1}^{m} \frac{e^{x_{(i)+}\beta}}{\left[\sum_{j\in R(t_{(i)})} e^{x_j\beta}\right]^{d_i}}, \tag{6.18}$$

gdzie d_i oznacza liczbę obiektów z czasem przeżycia $t_{(i)}$, zaś $x_{(i)+}$ jest równe sumie zmiennych po obiektach d_i, tj.

$$x_{(i)+} = \sum_{j\in D(t_{(i)})} x_j,$$

gdzie $D(t_{(i)})$ jest zbiorem badanych, u których czas życia wynosi $t_{(i)}$.

6.7.2. Aproksymacja Efrona. Jest nieco bardziej skomplikowana, ale też bliższa rzeczywistości. Funkcję częściowej wiarygodności przybliża się wyrażeniem:

$$l_{p2}(\beta) = \prod_{i=1}^{m} \frac{e^{x_{(i)+}\beta}}{\prod_{k=1}^{d_i}\left[\sum_{j\in R(t_{(i)})} e^{x_j\beta} - \frac{k-1}{d_i}\sum_{j\in D(t_{(i)})} e^{x_j\beta}\right]} \tag{6.19}$$

Gdy $d_i = 1$, wyrażenia w mianowniku we wszystkich trzech wzorach (podstawowym i obu przybliżonych) są sobie równe.

Wszystkie dalsze obliczenia — estymatora β i wariancji — przebiegają podobnie jak w prostszym przypadku.

P r z y k ł a d 6.2 (Badanie HMO (*Health maintenance organization*) pacjentów zakażonych wirusem HIV, [6]). W badaniu brało udział 100 pacjentów, o 31 różnych czasach przeżycia. Liczba osób z tym samym czasem przeżycia wahała się między 1 a 17. Wyniki estymacji poszczególnymi metodami przedstawia tabelka 6.3.

	WIEK		LEK	
Metoda	Współczynnik	Błąd stand.	Współczynnik	Błąd stand.
Dokładna	0.0977	0.0187	1.0226	0.2572
Breslow	0.0915	0.0185	0.9414	0.2555
Efron	0.0971	0.0186	1.0167	0.2562

Tabela 6.3. Porównanie wyników estymacji poszczególnymi metodami.

Potwierdzają one, że estymacja Efrona dostarcza wyników bliższych wynikom dokładnym. Tak naprawdę estymatory wyznaczone wszystkimi trzema metodami są sobie bliskie, a ich błędy standardowe są niemal identyczne, więc używając aproksymacji Breslowa powinniśmy otrzymać te same wnioski.

6.8. *Estymacja funkcji przeżycia.* W modelu Coxa funkcję przeżycia wyrażaliśmy np. wzorem

$$S(t, \boldsymbol{x}, \beta) = [S_0(t)]^{\exp \boldsymbol{x}'\beta}, \tag{6.20}$$

gdzie $S_0(t)$ oznaczało bazową funkcję przeżycia. Korzystając z powyższej zależności, przy znajomości estymatora współczynnika regresji, do estymacji funkcji przeżycia potrzeba nam już tylko estymatora bazowej funkcji przeżycia.

Aby go otrzymać, korzystać będziemy z estymatora Kaplana-Meiera. Wprowadzamy oznaczenie

$$\widehat{\alpha}_i = 1 - \frac{d_i}{n_i},$$

gdzie n_i jest liczbą obiektów będących w ryzyku zdarzenia w chwili $t_{(i)}$, zaś d_i — zaobserwowaną liczbą zdarzeń w tej chwili. Jest to estymator warunkowego prawdopodobieństwa przetrwania w chwili $t_{(i)}$ (indeksy w nawiasach oznaczają, że czasy są uporządkowane).

Estymator Kaplana-Meiera funkcji przetrwania jest iloczynem estymatorów dla indywidualnych prawdopodobieństw warunkowych.

Wprowadzamy oznaczenie na warunkowe prawdopodobieństwo bazowe przetrwania:

$$\alpha_i = \frac{S_0(t_{(i)})}{S_0(t_{(i-1)})},$$

prawdopodobieństwo warunkowe można wówczas wyrazić przez

$$\frac{S(t_{(i)}, x, \beta)}{S(t_{(i-1)}, x, \beta)} = \left\{ \frac{[S_0(t_{(i)})]^{\exp x'\beta}}{[S_0(t_{(i-1)})]^{\exp x'\beta}} \right\} = \left(\frac{S_0(t_{(i)})}{S_0(t_{(i-1)})} \right)^{\exp x'\beta} = \alpha_i^{\exp x'\beta}.$$

Oznaczamy $\widehat{\theta}_l = \exp x'\beta$. Estymator warunkowego bazowego prawdopodobieństwa przetrwania otrzymujemy rozwiązując równanie

$$\sum_{l \in D_i} \frac{\widehat{\theta}_l}{1 - \alpha_i^{\widehat{\theta}_l}} = \sum_{l \in R_i} \widehat{\theta}_l, \tag{6.21}$$

gdzie R_i oznacza obiekty w zbiorze ryzyka w chwili $t_{(i)}$ (czasy uporządkowane) a D_i — obiekty w zbiorze ryzyka z czasem przetrwania równym $t_{(i)}$.

Jeśli nie mamy danych niedokładnych, zbiór D_i zawiera dokładnie jeden obiekt i rozwiązaniem powyższego równania jest

$$\widehat{\alpha_i} = \left[1 - \frac{\widehat{\theta_i}}{\sum_{l \in R_i} \widehat{\theta_i}}\right]^{\widehat{\theta_i}^{-1}}. \tag{6.22}$$

Przy danych niedokładnych zaś rozwiązanie uzyskujemy iteracyjnie. Estymatorem bazowej funkcji przetrwania jest iloczyn indywidualnych estymatorów bazowych prawdopodobieństw przetrwania

$$\widehat{S}_0(t) = \prod_{t_{(i)} \leqslant t} \widehat{\alpha}_i. \tag{6.23}$$

Estymator funkcji przetrwania ($S(t, x, \beta)$) uzyskujemy podstawiając do wzoru (6.20) bazową funkcję przetrwania i estymatory współczynników β.

Niektóre pakiety, jako estymator bazowej funkcji przetrwania, przyjmują prostą funkcję estymatora warunkowego prawdopodobieństwa przetrwania, tj.

$$\widehat{h}_0(t_{(i)}) = 1 - \widehat{\alpha}_i.$$

Takie estymatory mogą często być niestabilne lub za mało gładkie, by ich używać. (Można jednak użyć pewnych metod ich wygładzania).

Estymator skumulowanej bazowej funkcji hazardu jest nieco bardziej praktyczny niż bazowej funkcji hazardu. Otrzymujemy go wykorzystując zależność

$$\widehat{S}_0(t) = e^{-\widehat{H}_0(t)},$$

stąd estymator skumulowanej bazowej funkcji hazardu to

$$\widehat{H}_0(t) = -\ln[\hat{S}_0(t)].$$

Dla konkretnych wartości współczynników mamy więc

$$\widehat{H_0}(t, x, \beta) = -\ln[\widehat{S_0}(t, x, \hat{\beta})] = -e^{x'\hat{\beta}} \ln[\widehat{S_0}(t)], \tag{6.24}$$

co, jako funkcja czasu, może dostarczyć użytecznego graficznego opisu ryzyka.

7. Model ze zmiennymi zależnymi od czasu.

7.1. *Zmienne zależne od czasu.* Z punktu widzenia modelu jego uogólnienie, aby zawierał w sobie ewentualną zależność zmiennych objaśniających od czasu, nie jest trudne, jednak znacznie go komplikuje. Należy więc ostrożnie i z umiarem uwzględniać takie zależności oraz analizować charakter takich zmiennych przed włączeniem ich do modelu.

Wartość zmiennej zależnej od czasu może zależeć jedynie od czasu trwania badania, niekoniecznie od czasu kalendarzowego. Należy zwrócić uwagę

na definicję takiej zmiennej w chwili zero, gdyż dla różnych badanych chwila ta może nastąpić w innym czasie kalendarzowym.

Zmienne zależne od czasu na ogół możemy podzielić na *wewnętrzne* i *zewnętrzne*. Zmienna wewnętrzna to taka, której wartość jest zależna od badanego obiektu i wymaga, by był on pod okresową obserwacją. Przykładem może być nowa terapia w leczeniu raka, z punktem końcowym będącym śmiercią z powodu nowotworu. Jeśli aktualna wartość pomiaru stanu fizjologicznego jest związana z progresją choroby, mierzenie tej zmiennej zależy od ciągłej obserwacji pacjenta.

Zmienna zewnętrzna natomiast to taka, której wartość w poszczególnej chwili nie wymaga bezpośredniej obserwacji badanego obiektu. Są to na przykład czynniki zależne od badania lub od środowiska, i wpływają one na wszystkich będących pod obserwacją. Przykładami takich zmiennych może być np. wiek pacjenta. Jeśli obserwujemy go przez dość długi okres czasu, jego aktualny wiek może mieć większy wpływ na przetrwanie niż na początku badania. (Oczywiście, gdy znamy datę urodzenia pacjenta, aktualny wiek może być wyliczony w każdej chwili, niezależnie od tego, czy pacjent znajduje się wciąż pod obserwacją). Inną ważną zewnętrzną zmienną zależną od czasu jest czas sam w sobie.

Aby zmienne zależne od czasu zostały uwzględnione w modelu, konieczna jest zmiana zapisu ([6]). Niech $x(t)$ oznacza wartość zmiennej x w chwili t. (Zakładamy, że wszystkie obiekty dołączają do badania w chwili 0). Niech $x_l(t_i)$ oznacza wartość zmiennej x dla obiektu l w chwili t_i. Dla wielu zmiennych objaśniających ($x_{lk}(t_i)$ — wartość k-tej zmiennej itd.) dostajemy więc wektor zmiennych:

$$x_l'(t_i) = [x_{l1}(t_i), \ldots, x_{lp}(t_i)].$$

Ten zapis jest całkiem ogólny, tzn. jeśli mamy w modelu zmienną niezależną od czasu, przyjmujemy $x_{lk}(t_i) = x_{lk}(t = 0) = x_{lk}$.

Uogólniony model hazardu przedstawia się wzorem:

$$h(t, x(t), \beta) = h_0(t) \exp[x'(t)\beta], \tag{7.1}$$

uogólniona funkcja częściowej wiarygodności zaś:

$$l_p(\beta) = \prod_{i=1}^{n} \left[\frac{e^{x_i'(t_{(i)})\beta}}{\sum_{l \in R(t_{(i)})} e^{x_l'(t_{(i)})\beta}} \right]. \tag{7.2}$$

Ważnym założeniem poczynionym w tym modelu jest, że współczynniki β nie zmieniają sie w czasie.

Estymatory otrzymuje się podobnie jak wcześniej.

P r z y k ł a d 7.1 (Badanie skuteczności terapii osób nadużywających leki (UIS), [6]). Badanie obejmowało dwa jednocześnie stosowane rodzaje terapii. Jego celem było porównanie skuteczności programów o różnym czasie

trwania. Czas przeżycia określamy jako czas, który upłynął od rozpoczęcia terapii do momentu powrotu do nałogu. Czas leczenia, oznaczany przez LOT (*length of treatment*) będziemy traktować jako zmienną zależną od czasu. Rozważamy możliwość, iż efekt leczenia zależy od czasu, przez który badany będzie brał w nim udział — mimo, że każdy mógł wycofać się z niego w dowolnej chwili (i wrócić do nałogu), ci z dłuższym stażem przejawiali tendencję do pozostawania w nim jeszcze dłużej (i również po zakończeniu dłużej nie wracali do nałogu).

W celu uwzględnienia takiej ewentualności w modelu definiujemy zmienną wskaźnikową zależną od czasu:

$$\text{OFF\_TRT}(t) = \begin{cases} 0 & \text{jeżeli } t \leqslant \text{LOT}, \\ 1 & \text{jeżeli } t > \text{LOT}. \end{cases}$$

Pozostałe oznaczenia (w nawiasach jednostki bądź dopuszczalne wartości):

AGE — wiek (lata)

BECKTOTA — ilość punktów w skali depresji Becka przy przyjęciu (0.000-54.000)

NDRUGFPx — liczba wcześniejszych terapii lekiem x (0-40)

IVHX_3 — wcześniejsze przyjmowanie leków drogą dożylną (1-nigdy, 2-wcześniej, 3-ostatnio)

RACE — kolor skóry (0-biały, 1-inny)

TREAT — randomizowany wybór metody leczenia (0-krótkie, 1-długie)

SITE — miejsce leczenia (0 = A, 1 = B)

Współczynniki wyestymowane w modelu bez zmiennej zależnej od czasu OFF_TRT(t) przedstawia tabela 7.1.

Zmienna	Współczynnik	Błąd stand.
AGE	−0.041	0.010
BECKTOTA	0.009	0.005
NDRUGFP1	−0.574	0.125
NDRUGFP2	−0.215	0.049
IVHX_3	0.228	0.109
RACE	−0.467	0.135
TREAT	−0.247	0.094
SITE	−1.317	0.531
AGExSITE	0.032	0.016
RACExSITE	0.850	0.248

Tabela 7.1. Współczynniki wyestymowane bez zmiennej udziału zmiennej zależnej od czasu.

Po uwzględnieniu zmiennej OFF_TRT(t) i ponownej estymacji okazuje się, że zarówno współczynnik dla tej zmiennej jest wysoki, jak i współczynnik

dla TREAT uległ znacznej zmianie. Inne wartości przyjmują też zmienne zawierające SITE (tabela 7.2).

Zmienna	Współczynnik	Błąd stand.
AGE	−0.038	0.010
BECKTOTA	0.008	0.005
NDRUGFP1	−0.609	0.128
NDRUGFP2	−0.226	0.050
IVHX_3	0.275	0.109
RACE	−0.517	0.135
TREAT	0.019	0.096
SITE	−0.969	0.516
AGExSITE	0.036	0.016
RACExSITE	0.511	0.257
OFF_TRT	2.571	0.157

Tabela 7.2. Współczynniki wyestymowane po uwzględnieniu zmiennej zależnej od czasu.

Współczynnik hazardu porównujący pacjenta po leczeniu do pacjenta w trakcie leczenia wynosi

$$\widehat{HR}(t) = e^{2.571} = 13.08,$$

co wskazuje, że osoby nie biorące aktywnego udziału w terapii są znacznie bardziej narażone na powrót do nałogu, niezależnie od metody i miejsca leczenia.

8. Model z danymi niepełnymi.

8.1. *Modyfikacja modelu.* Dotychczasowe rozważania brały pod uwagę dane cenzorowane jedynie prawostronnie, zatem każda obserwacja zaczynała się w chwili 0 i trwała aż do wystąpienia zdarzenia, zakończenia badania bądź wycofania się pacjenta.

W praktyce często spotykamy się z innymi rodzajami niepełnych danych, zarówno cenzorowanych, jak i obciętych; i lewo- (na początku obserwacji) i prawostronnie (na końcu obserwacji). Rzadko zdarza się jednoczesne cenzorowanie i obcięcie, spotykamy się jednak z obserwacjami niekompletnymi z obu stron (np. lewostronnie obciętymi i równocześnie prawostronnie cenzorowanymi).

Aby uwzględnić dane cenzorowane w modelu Coxa, potrzebne będą pewne nowe oznaczenia ([6]).

O i-tym obiekcie wiemy, że czas jego obserwacji T jest ograniczony dwiema wartościami ($a_i < T \leqslant b_i$). Ponadto dysponujemy informacją, czy

interesujące nas zdarzenie zaszło ($c_i = 1$ jeśli zaszło, $c_i = 0$, jeśli nie zaszło). I tak dla lewostronnego cenzorowania mamy

$$a_i = 0, c_i = 1,$$

dla prawostronnego natomiast

$$b_i = \infty, c_i = 0.$$

Prawdopodobieństwo wpadnięcia do przedziału i-tego obiektu wynosi:

$$[S(a_i, x_i, \beta)]^{1-c_i} \left[S(a_i, x_i, \beta) - S(b_i, x_i, \beta)\right]^{c_i}. \tag{8.1}$$

Upraszcza się to do:

- $1 - S(b_i, x_i, \beta)$
  dla cenzorowania lewostronnego,
- $S(a_i, x_i, \beta) - S(b_i, x_i, \beta)$
  dla danych niecenzorowanych,
- $S(a_i, x_i, \beta)$
  dla cenzorowania prawostronnego.

W funkcji wiarygodności będziemy mieli zatem iloczyn tych prawdopodobieństw dla wszystkich obiektów:

$$l(\beta) = \prod_{i=1}^{n} [S(a_i, x_i, \beta)]^{1-c_i} [S(a_i, x_i, \beta) - S(b_i, x_i, \beta)]^{c_i}. \tag{8.2}$$

Procedura dopasowywania modelu upraszcza się, gdy a i b przyjmują tylko kilka wartości — można wtedy odnieść się do wartości cenzorowanych przez przedziały czasowe jednakowe dla wszystkich badanych.

Przyjmijmy więc, że mamy $J + 1$ przedziałów $(t_{j-1}, t_j]$ dla $j = 1, 2, \ldots, J + 1$, $t_0 = 0$ i $t_{j+1} = \infty$ i że są one wspólne dla wszystkich obiektów. Dla wygody zapisu oznaczmy przez I_j j-ty przedział czasowy $(t_{j-1}, t_j]$. Zmienna wskaźnikowa dla określonego przedziału czasowego dla i-tego obiektu definiuje się następująco:

$$y_{ij} = \begin{cases} 1 & (a_i, b_i] = I_j \\ 0 & \text{wpp} \end{cases}.$$

Po przeformułowaniu prawdopodobieństwo dla j-tego przedziału będzie wyrażało się wzorem:

$$S(t_{j-1}, x_i, \beta) - S(t_j, x_i, \beta) = S(t_{j-1}, x_i, \beta) \left[1 - \frac{S(t_j, x_i, \beta)}{S(t_{j-1}, x_i, \beta)}\right]. \tag{8.3}$$

Wyrażenie w nawiasach kwadratowych określa prawdopodobieństwo, że zdarzenie miało miejsce w j-tym przedziale przy wiedzy, że obiekt żył pod koniec przedziału $j - 1$ (czyli $Pr(t_{j-1} < T \leqslant t_j | T > t_{j-1})$).

W modelu hazardów proporcjonalnych stosunek funkcji przeżycia na kolejnych końcach przedziałów może zostać zapisany jako:

$$\frac{S(t_j, x_i, \beta)}{S(t_{j-1}, x_i, \beta)} = \exp[-\exp(x_i'\beta + \tau_j)], \tag{8.4}$$

gdzie

$$\tau_j = \ln\left\{-\ln\left[\frac{S_0(t_j)}{S_0(t_{j-1})}\right]\right\}.$$

Na mocy wzoru (8.4) warunkowe prawdopodobieństwo zapisujemy jako

$$\theta_{ij} = 1 - \exp[-\exp(x_i'\beta + \tau_j)].$$

Funkcja wiarygodności może teraz zostać wyrażona przez

$$\begin{aligned} l(\beta) &= \prod_{i=1}^{n}\prod_{j=1}^{J+1}\left\{[S(t_{j-1}, x_i, \beta)^{1-c_i}][S(t_{j-1}, x_i, \beta) - S(t_j, x_i, \beta)]^{c_i}\right\}^{y_{ij}} \\ &= \prod_{i=1}^{n}\prod_{j=1}^{J+1}\left\{[S(t_{j-1}, x_i, \beta)]\left[1 - \frac{S(t_j, x_i, \beta)}{S(t_{j-1}, x_i, \beta)}\right]^{c_i}\right\}^{y_{ij}} \\ &= \prod_{i=1}^{n}\prod_{j=1}^{J+1}\{S(t_{j-1}, x_i, \beta) \times \theta_{ij}^{c_i}\}^{y_{ij}}. \end{aligned} \tag{8.5}$$

Funkcję przetrwania na końcu danego przedziału wyrazić można jako iloczyn kolejnych warunkowych prawdopodobieństw przetrwania. Po przekształceniach algebraicznych otrzymujemy, że:

$$S(t_{j-1}, x_i, \beta) = \prod_{l=1}^{j-1}(1 - \theta_{il}).$$

Podstawiając to do wzoru (8.5), otrzymujemy funkcję wiarygodności:

$$l(\beta) = \prod_{i=1}^{n}\prod_{j=1}^{J+1}\left[\prod_{l=1}^{j-1}(1 - \theta_{il})\theta_{ij}^{c_i}\right]^{y_{ij}}.$$

Oznaczmy teraz przez I_{k_i} przedział obserwowany dla i-tego obiektu, tzn. $I_{k_i} = (a_i, b_i]$. W powyższym wzorze na funkcję wiarygodności można zauważyć, że jedynym przypadkiem, gdy czynniki iloczynu po j różnią się od 1 jest gdy $j = k_i$. Stąd

$$l(\beta) = \prod_{i=1}^{n}\theta_{ik_i}^{c_i}\prod_{j=1}^{k_i-1}(1 - \theta_{ij}).$$

Powyższa funkcja może zostać przekształcona tak, aby przypominała funkcję wiarygodności dla binarnego modelu regresji. Definiujemy pseudo-

binarną zmienną wynikową $z_{ij} = y_{ij} \times c_i$ i teraz

$$l(\beta) = \prod_{i=1}^{n} \prod_{j=1}^{k_i - 1 + c_i} (1 - \theta_{ij})^{1 - z_{ij}} \theta_{ij}^{z_{ij}}. \tag{8.6}$$

Dla każdego obiektu i jest to wiarygodność dla $k_i - 1 + c_i$ niezależnych binarnych obserwacji z prawdopodobieństwami θ_{ij} i wynikami z_{ij}.

P r z y k ł a d 5.1 (Badanie skuteczności terapii osób nadużywających leki (UIS), model z danymi cenzorowanymi, [6]). Zakładamy, że czas badania jest rejestrowany co 6 miesięcy. Czas przetrwania jest więc określony przez 6-miesięczny przedział czasowy, w którym nastąpił powrót do nałogu lub ostatni, w którym wiemy, że pacjent nie nadużywał leków.

W badaniu mamy 7 przedziałów: $I_j = (6(j-1), 6j)$ dla $j = 1, ..., 6$ i $I_7 = (36, \infty]$. Wzięło w nim udział 628 pacjentów, przy czym w przedziale I_7 jest tylko jeden badany, w I_6 nie ma żadnego, zaś w I_5 jest ich czterech. Na dodatek obserwacje wszystkich tych pięciu osób są cenzorowane. Żeby natomiast móc estymować τ_j, każdy przedział musi zawierać przynajmniej jeden obiekt z czasem życia niecenzorowanym. Dlatego też dane z przedziałów I_4, I_5, I_6 i I_7 sumujemy.

Przykładowe dane po podziale na przedziały przedstawia tabelka 8.1.

ID	Miesiąc	Przedział	Cenzorowanie	z	Wiek
2	6	1	1	1	30
4	6	1	1	1	29
1	12	1	1	0	36
1	12	2	1	1	36
3	12	1	1	0	30
3	12	2	1	1	30
7	18	1	1	0	36
7	18	2	1	0	36
7	18	3	1	1	36
31	18	1	0	0	36
31	18	2	0	0	36
31	18	3	0	0	36
5	24	1	0	0	21
5	24	2	0	0	21
5	24	3	0	0	21
5	24	4	0	0	21
388	24	1	1	0	40
388	24	2	1	0	40
388	24	3	1	0	40
388	24	4	1	1	40

Tabela 8.1. Przykładowe dane po pogrupowaniu.

Pierwszy blok danych zawiera informacje o pacjentach, którzy powrócili do nałogu w czasie pierwszych 6 miesięcy.

Drugi zawiera po dwie linijki na badanego — pierwsza mówi o tym, że nie zaczął on z powrotem nadużywać leków w czasie pierwszych 6 miesięcy, druga — że zaczął w czasie kolejnych 6 ($z = 1$).

Kolejny blok składa się z 6 linii, po 3 na pacjenta. Pierwszy z nich (o $ID = 7$) wrócił do nałogu między 12. a 18. miesiącem (dla przedziału nr 3 ma on $z = 1$), zaś z drugim ($ID = 31$) ostatni kontakt miał miejsce w 18. miesiącu (dlatego dla niego $z = 0$, gdyż nie zaobserwowaliśmy powrotu do nałogu).

Podobnie w ostatnim bloku mamy dane o dwóch pacjentach, z których obserwacja jednego (o $ID = 5$) jest cenzorowana, zaś drugiego ($ID = 338$) nie. Wyestymowane parametry przedstawia tabelka 8.2.

Zmienna	Współczynnik	Błąd stand.
AGE	−0.040	0.010
BECKTOTA	0.006	0.005
NDRUGFP1	−0.531	0.130
NDRUGFP2	−0.197	0.050
IVHX_3	0.236	0.111
RACE	−0.444	0.137
TREAT	−0.235	0.097
SITE	−1.220	0.548
AGExSITE	0.029	0.017
RACExSITE	0.825	0.255
INT_1	1.827	0.432
INT_2	1.745	0.446
INT_3	0.904	0.475
INT_4	−0.816	0.728

Tabela 8.2. Wyestymowane wartości parametrów.

Parametry te nie odbiegają znacznie od wyestymowanych przy użyciu danych dokładnych, znaczy to, że możliwe jest uzyskanie dość dobrych estymatorów mimo danych pogrupowanych w przedziały.

8.2. *Estymacja funkcji przeżycia.* Estymując funkcję przeżycia zaczynamy od estymacji bazowej funkcji przeżycia S_0. Interesują nas jej wartości na końcach każdego z rozważanych przedziałów.

Z definicji τ_j oraz z tego, że $S_0(t_0 = 0) = 1$ na końcu pierwszego przedziału estymator bazowej funkcji hazardu wynosi:

$$\widehat{S_0}(t_1) = \exp[-\exp(\hat{\tau}_1)],$$

na końcu drugiego przedziału:

$$\widehat{S_0}(t_2) = \widehat{S_0}(t_1)\exp[-\exp(\hat{\tau}_2)]$$

i ogólnie, na końcu j-tego ($j = 1, 2, \ldots, J$) przedziału:

$$\widehat{S_0}(t_j) = \widehat{S_0}(t_{j-1})\exp[-\exp(\hat{\tau}_j)]. \tag{8.7}$$

Podobnie jak wcześniej, estymator funkcji przeżycia otrzymujemy z estymatora bazowej funkcji przeżycia (przy danym estymatorze współczynników regresji):

$$\hat{S}(t_j, x, \hat{\beta}) = [\widehat{S_0}(t_j)]^{\exp(x'\hat{\beta})}.$$

9. Podsumowanie. W niniejszej pracy przedstawiony został model hazardów proporcjonalnych Coxa oraz estymatory jego parametrów przy różnych założeniach co do zmiennych niezależnych oraz co do czasów przeżycia konkretnych obiektów. Do jego niewątpliwych zalet należą stosunkowo niewielkie ograniczenia czy dodatkowe założenia w jego wykorzystywaniu, dlatego też ma on szerokie zastosowanie. Szczególnie istotną możliwością, jaką daje jego stosowanie, jest uwzględnienie danych niepełnych, które w analizie przeżycia zdarzają się dość często.

Literatura

[1] Bender, Augustin, Blettner *Generating Survival Times to Simulate Cox Proportional Hazards Models* [http://epub.ub.uni-muenchen.de/1716/1/paper_338.pdf].

[2] Cox, D.R. *Regression Models and Life-Tables*, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 34, No. 2. (1972), 187-220 [http://www.stat.rutgers.edu/ rebecka/Stat687/cox.pdf].

[3] Elandt-Johnson R.C., Norman L. Johnson *Survival Models and Data Analysis*, John Wiley and Sons (1979).

[4] Fox J. *Cox Proportional-Hazards Regression for Survival Data* (2002) [http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-cox-regression.pdf].

[5] Fox J. *Introduction to Survival Analysis* (2006) [http://socserv.mcmaster.ca/jfox/Courses/soc761/survival-analysis.pdf].

[6] Hosmer, D.W., Jr, Lemeshow S. *Applied Survival Analysis. Regression Modeling of Time to Event Data*, John Wiley and Sons (1999).

[7] León, R.V., *Cox's Semiparametric Proportional Hazard rates Model: Motivating Example* (2000) [http://web.utk.edu/~leon/rel/class/Class2000/Cox.PDF].

[8] Ware, J.H., DeMets, D.L. (1976). *Reanalysis of some baboon descent data. Biometrics*, 32:459-463.

[9] Dane użyte w przykładach: ftp://ftp.wiley.com/public/sci_tech_med/survival/.

Agnieszka Deszyńska
Instytut Matematyki UJ
Kraków, Poland
E-mail: Agnieszka.Deszynska@im.uj.edu.pl

Models of Cox's Proportional Hazards

Abstract. The paper presents Cox proportional hazards model, its properties and methods of its parameters estimation. It is widely applicable in survival analysis – in prediction of survival chances of some objects (usually patients in medical studies). The essential advantage of the model is allowing of incomplete data, which often appear in studies – both in random and fixed way. Cox model works especially well when determination of treatment effectiveness in comparative sense (with reference to other therapies) is needed. Terminology and examples are usually taken from medicine but the model can be used also in sociology, crime detection or engineering.
Keywords: Cox model, hazard, survival analysis

(*wpłynęło 23 lipca 2010 r.*)