

**Uniwersytet Warszawski**  
Wydział Matematyki, Informatyki i Mechaniki

**Magda Młynarczyk**

Nr albumu: 339340

**Tu będzie tytuł**

**Praca magisterska**  
**na kierunku MATEMATYKA**  
**w zakresie MATEMATYKI STOSOWANEJ**

Praca wykonana pod kierunkiem  
**dra hab. Przemysława Biecka**  
Instytut Matematyki Stosowanej i Mechaniki

Maj 2017

## **Oświadczenie kierującego pracą**

Potwierdzam, że niniejsza praca została przygotowana pod moim kierunkiem i kwalifikuje się do przedstawienia jej w postępowaniu o nadanie tytułu zawodowego.

Data

Podpis kierującego pracą

## **Oświadczenie autora (autorów) pracy**

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Data

Podpis autora pracy

### **Streszczenie**

Tu będzie streszczenie

### **Słowa kluczowe**

analiza przeżycia, model ryzyk konkurujących

### **Dziedzina pracy (kody wg programu Socrates-Erasmus)**

11.2 Statystyka

### **Klasyfikacja tematyczna**

D. Software

D.127. Blabalgorithms

D.127.6. Numerical blabalysis



# Spis treści

<b>Wprowadzenie</b>	5
<b>1. Analiza przeżycia - teoria</b>	7
1.1. Podstawy	7
1.2. Podstawowe modele parametryczne	9
1.3. Estymacja modeli nieparametrycznych	10
1.4. Porównywanie modeli analizy przeżycia	10
1.5. Model Coxa	13
1.6. Modele ryzyk konkurujących	13
1.7. Model Coxa dla ryzyk konkurujących	14
<b>2. Biblioteka jakas nazwa</b>	15
2.1. Wprowadzenie	15
2.2. Algorytm	15
2.3. Opis dostępnych funkcji	15
<b>3. Przykład zastosowania na danych pacjentów z ostrą białaczką szpikową</b>	17
3.1. Opis danych	17
3.2. Eksploracja danych	17
3.3. Model ryzyk konkurujących	17
3.4. Zastosowanie mojej biblioteki	17
<b>4. Podsumowanie i dyskusja</b>	19
<b>5. Kod źródłowy</b>	21
5.1. użyty kod	21
<b>Bibliografia</b>	23



# Wprowadzenie

Tu będzie wprowadzenie.





# Rozdział 1

## Analiza przeżycia - teoria

### 1.1. Podstawy

Analiza przeżycia jest gałęzią statystyki zajmującą się badaniem czasu do wystąpienia danego zdarzenia oraz czynników wpływających na ten czas. Pojęcie zdarzenia obejmuje szerokie spektrum wydarzeń i zjawisk, takie jak śmierć, choroba, niewypłacalność czy awaria urządzenia. Dzięki temu, metody jakie oferuje nam analiza przeżycia mogą być stosowane w bardzo wielu dziedzinach. Pojęcie zdarzenia określa się czasem jako 'porażka', mimo, iż może ono odnosić się także do pozytywnych wydarzeń (np. poprawa stanu zdrowia pacjenta). Bardzo ważnym spostrzeżeniem jest fakt, że w standardowej analizie przeżycia rozważamy tylko jedno zdarzenie, które może wystąpić u każdej jednostki. W przypadku więcej niż jednego możliwego zdarzenia rozważamy modele zdarzeń rekurencyjnych bądź modele ryzyk konkurencyjnych, o których mowa będzie w następnych podrozdziałach.

Podstawowym celem analizy przeżycia, jest nie tylko modelowanie i interpretacja rozkładu czasu przeżycia w danej populacji. Bardzo istotnym elementem jest tutaj także porównywanie tych rozkładów w różnych grupach (np. w przypadku badania efektu placebo) oraz związek między tymi rozkładami a zmiennymi na nie wpływającymi - tak zwanymi zmiennymi objaśniającymi.

Fundamentalną częścią analizy przeżycia jest zdefiniowane jednoznacznie określonej zmiennej losowej  $T$  reprezentującej czas od określonego punktu w czasie do wystąpienia zdarzenia. Zmienna ta może być wyrażona w dowolnej jednostce czasu (sekundy, dni, lata...). Drugą niezbędną definicją jest określenie zmiennej  $d$ , oznaczającej, czy dana obserwacja była cenzorowana - to znaczy, że czas początku obserwacji bądź czas wystąpienia zdarzenia nie jest znany. Formalnie:

$$d_i = \begin{cases} 0 & \text{gdy } i - \text{ta obserwacja jest cenzorowana} \\ 1 & \text{w przeciwnym przypadku} \end{cases} \quad (1.1)$$

gdzie:

$i \in \{1, 2, \dots, N\}$  – numer obserwacji.

Najczęściej występującym przykładem cenzorowania jest cenzorowanie prawostronne, kiedy wiemy, że zdarzenie nie wystąpiło przed danym czasem  $T^*$ . Rzadziej występuje cenzorowanie lewostronne, kiedy wiemy, że zdarzenie wystąpiło przed danym czasem  $T^{**}$  oraz cenzorowanie przedziałowe kiedy wiemy, że wydarzenie wystąpiło w danym przedziale czasowym  $[T^{**}, T^*]$

Możemy wyróżnić trzy typy cenzorowania:

- typ I - przedział czasowy, w którym jednostki są obserwowane jest z góry określony

- typ II - jednostki są obserwowane, aż zdarzenie wystąpi w określonej frakcji badanych jednostek

- losowe

**POPRAW, sa str 37-39** oraz trzy założenia dotyczące cenzorowania:

- niezależność - ...
- losowość - ...
- **informatywność?** - ...

Aby określić rozkład przeżycia potrzebujemy następujących, kluczowych definicji:

**Definicja 1.1.1** *Funkcja przeżycia* - funkcja  $S : [0, \infty) \rightarrow [0, 1]$  dana wzorem:

$$S(t) = \mathbf{P}(t < T), \quad 0 < t < \infty$$

Funkcja przeżycia określa prawdopodobieństwo przeżycia do chwili  $t$ , dając nam najistotniejszą informację, jaką dostajemy z naszych danych dla analizy przeżycia. Jest ona niemalejącą, prawostronnie ciągłą funkcją czasu. Zachodzi:

$$S(0) = 1$$

Funkcję przeżycia często definiuje się także w terminach funkcji hazardu.

**Definicja 1.1.2** *Funkcja hazardu* - funkcja  $h : [0, \infty) \rightarrow \mathbf{R}$  dana wzorem:

$$h(t) = \lim_{\delta \rightarrow 0} \frac{\mathbf{P}(t < T < t + \delta | T > t)}{\delta}$$

Funkcja hazardu jest prawdopodobieństwem, że zdarzenie nastąpi w następnym 'dowolnie krótkim' przedziale czasu, jeżeli wiemy, że nie wystąpiło przed czasem  $t$ , podzielonym przez długość tego przedziału czasowego. Jest to funkcja nieujemna, nieograniczona z góry. Nazywana jest także funkcją ryzyka.

Zdefiniowane powyżej dwie funkcje pozwalają na określenie rozkładu przeżycia. Do dalszych analiz przydatnych jest jednak jeszcze kilka definicji.

**Definicja 1.1.3** *Dystrybuanta funkcji ryzyka* - funkcja  $F : [0, \infty) \rightarrow [0, 1]$  dana wzorem:

$$F(t) = \mathbf{P}(t \leq T)$$

**Definicja 1.1.4** *Gęstość prawdopodobieństwa* - funkcja  $f : [0, \infty) \rightarrow \mathbf{R}$  dana wzorem:

$$f(t) = -\frac{d}{dt}S(t) = \frac{d}{dt}F(t)$$

Na podstawie powyższych definicji otrzymujemy zależność

$$h(t) = \frac{f(t)}{S(t)} \tag{1.2}$$

To znaczy, że hazard w momencie  $t$  jest prawdopodobieństwem, że zdarzenie pojawi się w 'okolicach' momentu  $t$  podzielonym przez prawdopodobieństwo, że zdarzenie nie pojawiło się do czasu  $t$ .

**Definicja 1.1.5** *Dystrybuanta funkcji hazardu* - funkcja  $H : (0, \infty) \rightarrow \mathbf{R}$  dana wzorem:

$$H(t) = \int_0^t h(u) du$$

Dystrybuanta funkcji hazardu w punkcie  $t$  jest zdefiniowana jako pole pod wykresem funkcji hazardu do momentu  $t$ . Funkcję przeżycia możemy teraz zapisać w postaci:

$$S(t) = \exp\left(-\int_0^t h(u) du\right) = \exp(-H(t)) \quad (1.3)$$

## 1.2. Podstawowe modele parametryczne

W analizie przeżycia zakłada się czasami dany rozkład przeżycia, otrzymując model parametryczny. Najprostszym przykładem jest model eksponencjalny, w którym zakłada się stały hazard:

$$h(t) = \lambda \quad (1.4)$$

Wówczas otrzymujemy:

$$H(t) = \int_0^t h(u) du = \int_0^t \lambda du = \lambda t$$

$$S(t) = \exp(-H(t)) = \exp(-\lambda t)$$

$$f(t) = h(t)S(t) = \lambda \exp(-\lambda t)$$

Założenie stałego hazardu często nie jest spełnione. Innym często używanym modelem jest model o rozkładzie Weibulla z funkcją hazardu:

$$h(t) = \alpha \lambda (\lambda t)^{\alpha-1}, \quad \alpha, \lambda > 0 \quad (1.5)$$

Dla tego modelu otrzymujemy:

$$H(t) = \int_0^t h(u) du = \alpha \lambda^\alpha \int_0^t u^{\alpha-1} du = \alpha \lambda^\alpha \frac{1}{\alpha} u^\alpha \Big|_0^t = (\lambda t)^\alpha$$

$$S(t) = \exp(-H(t)) = \exp(-(\lambda t)^\alpha)$$

Rozkład eksponencjalny jest specjalnym przypadkiem rozkładu Weibulla dla parametru  $\alpha = 1$ . Dla  $\alpha > 1$  funkcja hazardu jest rosnąca, dla  $\alpha < 1$  jest malejąca.

Funkcję przeżycia estymuje się także za pomocą rozkładu lognormalnego. Mamy wówczas:

$$S(t) = 1 - \Phi\left(\frac{\log(t) - \mu}{\sigma}\right) \quad (1.6)$$

gdzie:

$\Phi$  - dystrybuanta rozkładu normalnego  $\mathbf{N}(0, 1)$  Funkcja hazardu w tym przypadku monotonicznie rośnie od 0 do swojego maksimum, a następnie monotonicznie maleje do 0 przy  $t \rightarrow 0$ . Dlatego model ten jest przydatny, kiedy prawdopodobieństwo wystąpienia zdarzenia rośnie na początku obserwacji, a później maleje.

Kolejnym rozkładem używanym do modelowania przeżycia jest rozkład gamma, o gęstości danej:

$$f(t) = \frac{\lambda^\beta t^{\beta-1} \exp(-\lambda t)}{\Gamma(\beta)}, \quad \lambda, \beta > 0 \quad (1.7)$$

Dla tego modelu funkcja przeżycia oraz funkcja hazardu nie dają zapisać się w prostej formie, mogą być jednak obliczone za pomocą wzorów z poprzedniego podrozdziału. Ponownie, Rozkład gamma, dla parametru  $\beta = 1$  sprowadza się do rozkładu wykładniczego. Dla  $\beta < 1$  funkcja hazardu jest rosnąca, dla  $\beta > 1$  jest malejąca.

### 1.3. Estymacja modeli nieparametrycznych

W wielu przypadkach nie jesteśmy w stanie wybrać odpowiedniej rodziny parametrycznej do opisu naszego modelu. Zajmujemy się wtedy modelami nieparametrycznymi, posiadającymi bardziej elastyczne właściwości modelowania.

Podstawowym estymatorem funkcji przeżycia używanym w analizie przeżycia jest **estymator Kaplana-Mayer’a**, dany wzorem:

$$\hat{S}(t) = \prod_{t_i \leq t} (1 - \hat{q}_i) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \quad (1.8)$$

gdzie:

$n_i$  - liczba jednostek narażonych na wystąpienie zdarzenia w czasie  $t_i$ ,

$d_i$  - liczba jednostek u których nastąpiło zdarzenie w czasie  $t_i$ .

Estymator Kaplana-Mayer’a jest nierosnącą funkcją schodkową, prawostronnie ciągłą. Najczęściej stosowanym estymatorem wariancji dla krzywej Kaplana-Mayera jest estymator zaproponowany przez Majora Greenwooda [3] w 1926 roku, dany wzorem:

$$\text{var}(\hat{S}(t)) \approx [\hat{S}(t)]^2 \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)} \quad (1.9)$$

Zauważmy, że

$$\forall_i \frac{d_i}{n_i(n_i - d_i)} = \frac{1}{n_i} P(T > t_i) P(T \geq t_i)$$

Przedział ufności na poziomie 95% wynosi wówczas:

$$\hat{CI}(t) = [-1.96\sqrt{\text{var}(\hat{S}(t))}, 1.96\sqrt{\text{var}(\hat{S}(t))}] \quad (1.10)$$

### 1.4. Porównywanie modeli analizy przeżycia

Głównym tematem mojej pracy jest porównywanie modeli ryzyk konkurujących pośród danych grup. Aby móc się tym zajmować, należy najpierw zrozumieć ideę porównywania modeli w przypadku zwykłej analizy przeżycia, którą będziemy później rozszerzać na przypadek ogólniejszy.

W przypadku modeli parametrycznych stosować można fundamentalne testy statystyczne, takie jak test t-studenta, jeżeli możemy założyć normalność rozkładu, bądź test Manna-Whitney’a, jeżeli założenie o normalności rozkładu nie jest spełnione. Jeżeli chcemy dopasować odpowiedni parametr danego rozkładu do naszych obserwacji, możemy użyć metodę największej wiarygodności.

Dla modeli nieparametrycznych potrzebujemy skonstruować test porównujący dwie krzywe przeżycia. Jako, że porównujemy ze sobą dwie krzywe, test statystyczny ze standardową hipotezą zerową i alternatywną:

$$H_0 : S_1(t) = S_0(t)$$

$$H_1 : S_1(t) \neq S_0(t)$$

nie jest adekwatny. Dwie krzywe przeżycia, mogą się krzyżować, albo być podobne na jednym odcinku oraz różne na innym odcinku czasu. W związku z tym, wprowadzamy rozwiązanie zwane **alternatywą Lehmana**:

$$H_1 : S_1(t) = [S_0(t)]^\psi$$

Równoważnie, dostajemy test hipotezy zerowej:

$$H_0 : \psi = 1 \quad (1.11)$$

przeciwko hipotezie alternatywnej:

$$H_1 : \psi < 1 \quad (1.12)$$

Oznacza to, że przy założeniu hipotezy alternatywnej, czasy przeżycia w grupie 1 będą dłuższe niż te w grupie 0. W analizie przeżycia, grupę 0 często traktuje się jako grupę kontrolną, a grupę 1 jako grupę testową.

Do skonstruowania testu, dla każdego czasu  $t_i$  potrzebujemy stworzyć tabelkę wielkości  $2 \times 2$  zawierającą liczbę jednostek u których nastąpiło zdarzenie i u których nie nastąpiło zdarzenie w czasie  $t_i$  dla obydwu grup.

Tablica 1.1: Tabela przeżycia w czasie  $t_i$

	Grupa kontrolna	Grupa testowa	Razem
Liczba zdarzeń	$d_{0i}$	$d_{1i}$	$d_i$
Liczba jednostek bez zdarzenia	$n_{0i} - d_{0i}$	$n_{1i} - d_{1i}$	$n_i - d_i$
Razem	$n_{0i}$	$n_{1i}$	$n_i$

Zakładając, że liczba zdarzeń w grupie kontrolnej i testowej jest niezależna, otrzymujemy hipergeometryczny rozkład  $d_{0i}$  pod warunkiem  $n_{0i}, n_i, d_i$ :

$$\mathbf{P}(d_{0i} | n_{0i}, n_{1i}, d_i) = \frac{\binom{n_{0i}}{d_{0i}} \binom{n_{1i}}{d_{1i}}}{\binom{n_i}{d_i}} \quad (1.13)$$

gdzie:

$$\binom{n}{d} = \frac{n!}{d!(n-d)!}, \quad n! = n * (n-1) * \dots * 2 * 1$$

Możemy teraz obliczyć średnią i wariancję zmiennej  $d_{0i}$ :

$$e_{0i} = \mathbf{E}d_{0i} = \frac{d_{0i}d_i}{n_i}$$

$$v_{0i} = \text{var}(d_{0i}) = \frac{n_{0i}n_{1i}d_i(n_i - d_i)}{n_i^2(n_i - 1)}$$

W następnym kroku sumujemy wszystkie różnice wartości obserwowanych i oczekiwanych zmiennej  $d_0$ , otrzymując liniową statystykę:

$$U_0 = \sum_{i=1}^N (d_{0i} - e_{0i})$$

$$V_0 = \text{var}(U_0) = \sum_{i=1}^N v_{0i}$$

Teraz możemy skonstruować statystykę testową:

$$\frac{U_0^2}{V_0} \sim \chi_1^2 \quad (1.14)$$

Powyższy test nazywany jest **logarytmicznym testem rang**. Test ten można uogólnić na tak zwany **ważony logarytmiczny test rang**, taki, że:

$$U_0(w) = \sum_{i=1}^N w_i (d_{0i} - e_{0i})$$

$$V_0(w) = \text{var}(U_0(w)) = \sum_{i=1}^N w_i^2 v_{0i}$$

Istnieje wiele testów opierających się na powyższej formule, zakładających różne postaci wag. Jednym z nich jest **test Wilcoxona**, dla którego wagą w czasie  $i$  jest liczba jednostek pod ryzykiem w czasie  $i$ :

$$w_i = n_i$$

**Test Tarone-Ware'a** przypisuje większą wagę do zdarzeń mających miejsce wcześniej, poprzez wykorzystanie pierwiastka z liczby jednostek pod ryzykiem jako wagi:

$$w_i = \sqrt{n_i}$$

**Test Flemminga-Harringtona** [4] daje największą elastyczność w wyborze statystyki testowej, poprzez wybranie parametru  $\rho$ :

$$w_i = N(\hat{S}(t_i))^\rho \quad (1.15)$$

W środowisku R w pakiecie **survival** [5] porównanie krzywych przeżycia za pomocą testu Flemminga-Harringtona może być wykonane za pomocą funkcji **survdif**.

Innym możliwym sposobem na porównanie modeli analizy przeżycia jest wykonanie **testu warstwowego** (ang. stratified test). Jest to kolejna modyfikacja logarytmicznego testu rangowego, używana w przypadku kiedy mamy kategorię zmienną objaśnianą  $G$  o niewielkiej liczbie poziomów  $G \in \{g_1, g_2, \dots, g_{n_g}\}$ . Zmienna  $G$  może oznaczać na przykład płeć, grupę wiekową czy podawaną dawkę leku. Testujemy wówczas hipotezę zerową:

$$H_0 : h_{0j}(t) = h_{1j}(t), \quad \text{dla } j = 1, 2, \dots, n_g$$

Dla każdej wartości zmiennej  $G = g$  obliczamy statystyki  $U_{0g}$  oraz  $V_{0g}$  a następnie wyznaczamy statystykę testową:

$$X^2 = \frac{(\sum_{n=1}^{n_g} U_{0g_n})^2}{\sum_{n=1}^{n_g} V_{0g_n}^2} \sim \chi_{n_g-1}^2$$

Statystyka testowa w powyższym teście różni się od tej ze logarytmicznego testu rang tym, że różnica zdarzeń obserwowanych i oczekiwanych jest sumowana po wszystkich czasach zdarzeń w każdej warstwie, a następnie różnice te są sumowane po wszystkich warstwach.

W pakiecie `survival` zaimplementowana została funkcja `strata` identyfikująca zmienne warstwowe, która może być wykorzystywana przy tworzeniu modeli proporcjonalnego hazardu.

## 1.5. Model Coxa

**Model Coxa**, nazywany także **modelem proporcjonalnego hazardu**, został po raz pierwszy zaproponowany przez Sir Davida Coxa i opiera się na **założeniu proporcjonalnego hazardu**:

$$h_1(t) = \Psi h_0(t) \quad (1.16)$$

które stwierdza, że zmienne objaśniane w modelu nie zależą od czasu i wpływają na funkcję hazardu w sposób multiplikatywny. Nazwa powyższego założenia odnosi się do faktu, że dla dwóch obserwacji iloraz ich funkcji hazardu jest stały. Model Coxa zakłada następującą postać funkcji hazardu:

$$h_1(t) = h_0(t)e^{\beta_1 z_1 + \beta_2 z_2 + \dots + \beta_m z_m} \quad (1.17)$$

gdzie:

$h_0(t)$  - hazard bazowy

$z_1, z_2, \dots, z_m$  - zmienne objaśniane, niezależne od czasu  $t$

$\beta_1, \beta_2, \dots, \beta_m$  - parametry

Powyższy model jest ogólną postacią regresji logistycznej.

W przypadku, gdy w modelu nie ma zmiennych objaśnianych, bądź wszystkie zmienne wynoszą 0, funkcja hazardu przyjmuje postać hazardu bazowego.

Ważną cechą modelu Coxa jest to, że postać funkcji  $h_0(t)$  jest nieokreślona (model jest **se-miparametryczny**), co czyni go adekwatnym do modelowania w różnych sytuacjach. Jeżeli nie jesteśmy pewni co poprawnego rozkładu naszych danych, użycie modelu Coxa powinno dać nam porównywalne wyniki do tych, które uzyskalibyśmy używając poprawnego rozkładu. Dodatkowo, mimo, iż nie znamy postaci hazardu bazowego, jesteśmy w stanie estymować parametry  $\beta_1, \dots, \beta_m$ . Dzięki tym własnościom jest on najczęściej używanym modelem w analizie przeżycia.

Estymację współczynników  $\beta_i$  można wykonać przy pomocy metody największej wiarygodności. Jako, że nie znamy postaci hazardu bazowego, korzystamy tutaj z częściowej wiarygodności:

$$L(\beta) = \prod_{i:d_i=1} \frac{e^{X_i \beta_i}}{\sum_{j:t_j > t_i} e^{X_1 \beta_1 + \dots + X_m \beta_m}} \quad (1.18)$$

## 1.6. Modele ryzyk konkurujących

Głównym zagadnieniem, którym zajmuję się w tej pracy są **modele ryzyk konkurujących**. Jest to jedno z dwóch podstawowych, obok modeli wielostanowych (ang. *multistate models*), uogólnień analizy przeżycia. W przypadku modeli ryzyk konkurujących mamy do czynienia z więcej niż jednym możliwym zdarzeniem i obserwujemy czas do wystąpienia pierwszego z nich, w odróżnieniu do modeli wielostanowych, w których, po wystąpieniu jednego zdarzenia może wystąpić następne, zgodnie z danym prawdopodobieństwem przejścia. Przykładem ryzyk

konkurujących może być np. zgon pacjenta z powodu raka płuc i zgon z innego powodu, czy wybór przez pracownika autobusu lub tramwaju jako środka transportu do pracy.

Formalnie, określamy nasz **zbiór ryzyk** jako  $\{1, 2, 3, \dots, K\}$

Jedną z możliwości modelowania ryzyk konkurujących jest używanie standardowych metod analizy przeżycia dla wszystkich ryzyk osobno, traktując inne zdarzenia jako cenzorowane. To podejście nie jest jednak poprawne w większości zastosowań, gdyż wymaga założenia o niezależności ryzyk, a nawet w tym przypadku interpretacja wyników bywa wątpliwa. Problemy powstające przy zastosowaniu takiego podejścia zaprezentuję na przykładzie w następnym rozdziale.

## 1.7. Model Coxa dla ryzyk konkurujących

Pierwszym opisanym sposobem na modelowanie ryzyk konkurujących będzie zastosowanie modeli Coxa dla każdego z ryzyk osobno, traktując wszystkie pozostałe obserwacje jako cenzorowane. W tym przypadku mamy:

**Definicja 1.7.1** *Funkcja hazardu dla ryzyka  $k$  - funkcja  $h_k : [0, \infty) \rightarrow \mathbf{R}$  dana wzorem:*

$$h_k(t) = \lim_{\delta \rightarrow 0} \frac{\mathbf{P}(t < T_k < t + \delta | T_k > t)}{\delta}, \quad k \in \{1, 2, 3, \dots, K\}$$

gdzie:  $T_k$  - czas do zdarzenia rodzaju  $k$ .

Model Coxa zastosowany do ryzyka  $k$  przyjmuje wówczas postać:

$$h_k(t) = h_{0,k}(t) e^{\beta_{1,k}z_1 + \beta_{2,k}z_2 + \dots + \beta_{m,k}z_m}, \quad k \in \{1, 2, 3, \dots, K\} \quad (1.19)$$

Takie podejście nie tylko zakłada niezależność ryzyk, ale także wymaga estymację  $K$  modeli, podczas gdy używając innych metod, wystarczy nam estymacja jednego modelu.



## Rozdział 2

# Biblioteka `jakas nazwa`

### 2.1. Wprowadzenie

możliwości zastosowanie, porównanie z innymi rozwiązaniami

### 2.2. Algorytm

Jak porównujemy grupy

### 2.3. Opis dostępnych funkcji

Opis wszystkich funkcji z pakietu



## Rozdział 3

# Przykład zastosowania na danych pacjentów z ostrą białaczką szpikową

### 3.1. Opis danych

Skąd mam dane, o co w nich chodzi

### 3.2. Eksploracja danych

tabelki, wykresiki

### 3.3. Model ryzyk konkurujących

Model ryzyk konkurujących zastosowany do moich danych

### 3.4. Zastosowanie mojej biblioteki

porównanie modeli ryzyk konkurujących w różnych grupach



## Rozdział 4

# Podsumowanie i dyskusja

co udało mi się zrobić i co możnaby ulepszyć



## Rozdział 5

# Kod źródłowy

### 5.1. użyty kod

Załącz kod?





# Bibliografia

- [1] Dirk F. Moore, *Applied Survival Analysis Using R*, Springer, 2016
- [2] David G. Kleinbaum, Mitchel Klein, *Survival Analysis*, Springer, 2012
- [3] S. Sawyer, *The Greenwood and Exponential Greenwood Confidence Intervals in Survival Analysis*, 2003
- [4] D.Cox, D.Oakes, *Analysis of Survival Data*, Chapman and Hall/CRC, London, New York, 1984
- [5] jak dodać manuale?? survival package manual
- [6] J Beyersmann, M.Schumacher, A. Allignol, *Competing Risks and Multistate Models with R*, Springer, 2012