Uniwersytet Warszawski
Wydział Matematyki, Informatyki i Mechaniki

Magda Młynarczyk
Nr albumu: 339340

Modele ryzyk konkurujących wraz z zastosowaniami w analizie chorych z nowotworami układu krwiotwórczego

Praca magisterska
na kierunku MATEMATYKA
w zakresie MATEMATYKI STOSOWANEJ

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

Czerwiec 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

Celem poniżej pracy jest zaprezentowanie narzędzi analizy danych z uwzględnieniem występowania ryzyk konkurujących. Przedstawione zostały metody i modele stosowane w analizie danych tego typu, jak również narzędzie, pakiet `cr17` dla programu R, który opracowałam w ramach pracy magisterskiej. Pakiet ten generuje raport zestawiający wykresy i tabele diagnostyczne oraz wyniki testów sprawdzających istotność występowania różnic w modelach. Praca składa się z trzech głównych części - podstaw teoretycznych, opisu struktury pakietu oraz funkcji w nim zaimplementowanych, a także przykładu jego zastosowania do danych medycznych o pacjentach z trzema podtypami nowotworów układu krwiotwórczego.

Słowa kluczowe

Analiza przeżycia, modele ryzyk konkurujących, porównywanie modeli statystycznych, analiza danych medycznych, model Coxa, funkcja skumulowanych częstości, R

Dziedzina pracy (kody wg programu Socrates-Erasmus)

11.1 Matematyka

Klasyfikacja tematyczna

62-07, 62P10, 62N03

Spis treści

Wprowadzenie

Podczas ostatniego roku studiów magisterskich brałam udział w projekcie o nazwie *InfAza*, pod kierownictwem dra n. med. Krzysztofa Mądrego oraz dra hab. Przemysława Biecka, realizowanym przez Uniwersytet Warszawski wraz z Warszawskim Uniwersytetem Medycznym. Projekt ten zajmował się badaniem wystąpień infekcji wśród pacjentów z nowotworami układu krwiotwórczego, poddanych terapii azacytadyną. Ważnym elementem efektywnego leczenia tego typu chorób jest poprawne wyłonienie chorych, którym należy podać profilaktykę przeciwwirusową, przeciwgrzybiczą lub przeciwbakteryjną, ponieważ znajdują się oni w grupie największego ryzyka zachorowalności na dany rodzaj infekcji.

Głównym celem tego projektu było więc znalezienie czynników wpływających na wystąpienie infekcji w ciągu trzech pierwszych miesięcy leczenia. Kolejnym etapem było badanie wystąpień infekcji w czasie, co doprowadziło nas do zagadnień analizy przeżycia. Po uwzględnieniu także zgonu jako możliwego zdarzenia zaczęliśmy zajmować się modelami ryzyk konkurujących. Szybko okazało się, że w `R` nie istnieje kompleksowa biblioteka pozwalająca na efektywną analizę naszych danych.

Stąd powstał pomysł na stworzenie, w ramach mojej pracy magisterskiej, biblioteki o nazwie `cr17`, umożliwiającej wygenerowanie sumarycznego raportu za pomocą tylko jednej funkcji. W raporcie tym znajdują się wykresy i tabelki diagnostyczne, oraz wyniki testów sprawdzających różnice pomiędzy modelami w poszczególnych grupach. Stworzenie tej biblioteki umożliwiło nam sprawne porównanie wielu modeli, w rozróżnieniu na różne podziały na grupy. Takie analizy są niezwykle ważne w medycynie, ponieważ pozwalają na określenie skuteczności różnych metod leczenia i pozwalają na skonstruowanie optymalnych schematów terapii.

Co istotne, zastosowania zarówno analizy przeżycia jak i ryzyk konkurujących są bardzo szerokie i nie ograniczają się jedynie do celów medycznych. Przykładem analizy, do której zastosowana może być biblioteka `cr17` jest badanie czasu działania maszyny do danego rodzaju awarii. Grupami pomiędzy którymi badać będziemy różnice, mogą być wtedy warunki, w jakich urządzenie było użytkowane (np. temperatura otoczenia). W ubezpieczeniach modelować możemy czas do wystąpienia jednego ze zdarzeń objętych polisą. Klientów można podzielić na grupy ze względu na wiek (co okazuje się być istotne w np. przypadku ubezpieczeń samochodowych). Przez ryzyko możemy rozumieć też zdarzenie pozytywne, na przykład zakup towaru danej kategorii w sklepie internetowym. Naturalnym podziałem klientów na grupy jest wtedy rozróżnienie ze względu na płeć.

Praca składa się z trzech rozdziałów. W pierwszym opisane jest podłoże metodologiczne, na podstawie którego zaimplementowane są funkcje w bibliotece `cr17`. Drugi rozdział stanowi prezentację pakietu. Znajduje się tam opis struktury pakietu, przegląd dostępnych funkcji oraz interpretacja końcowego raportu, jaki można uzyskać za pomocą głównej funkcji `summarizeCR` {`cr17`}. W trzecim rozdziale przedstawiam eksplorację danych z projektu *Infaza* oraz wyniki z zastosowania biblioteki do tych danych.

Rozdział 1

Analiza przeżycia i modele ryzyk konkurujących - teoria

1.1. Podstawy analizy przeżycia

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, takich jak śmierć, choroba, niewypłacalność kredytobiorcy 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. Na przykład, badany może być czas od początku podawania lekarstwa do poprawy stanu zdrowia, gdzie rozumiemy przez to uzyskanie wyniku danego badania krwi w przyjętych granicach. W 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 konkurujących, o których mowa będzie w następnych podrozdziałach.

Podstawowym celem analizy przeżycia jest modelowanie i interpretacja rozkładu czasu przeżycia w danej populacji. Istotnym elementem jest także porównywanie tych rozkładów w różnych grupach (np. w przypadku badania efektu placebo).

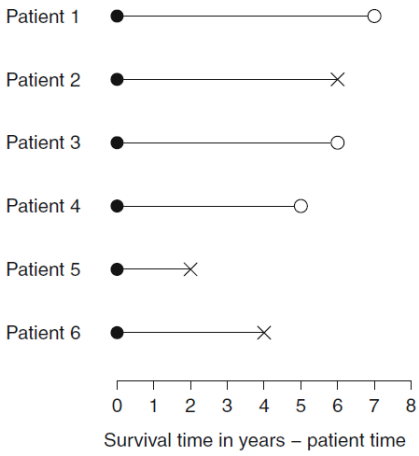
Fundamentalną częścią analizy przeżycia jest zdefiniowanie zmiennej losowej T, reprezentującej czas od określonego punktu w czasie do wystąpienia zdarzenia. Zmienna ta może być wyrażona w dowolnej jednostce czasu (sekundy, dni, lata...). Drugą niezbędną definicją jest określenie zmiennej losowej c, oznaczającej,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

$$c_i = \begin{cases} 0 & \text{gdy } i-ta \text{ obserwacja jest cenzorowna,} \\ 1 & \text{w przeciwnym przypadku,} \end{cases} \tag{1.1}$$

gdzie: $i \in \{1, 2, ..., N\}-$ numer obserwacji.

Kluczowym założeniem na temat zmiennej c jest to, że jest ona niezależna od wystąpień badanych zdarzeń.

Rodzajem cenzorowania, jakim będę zajmować się w tej pracy, jest cenzorowanie prawostronne, kiedy wiemy, że zdarzenie nie wystąpiło przed danym czasem T'. Takie dane można przedstawić graficznie albo w tabeli, jak zaprezentowano na rysunku 1.1 (ilustracja pochodzi z [13]).



Patient	Survtime	Status
1	7	0
2	6	1
3	6	0
4	5	0
5	2	1
6	4	1

Rysunek 1.1: Dane prawostronnie cenzorowane dotyczące czasu przeżycia wśród pacjentów. Na osi poziomej wykresu po lewej stronie znajduje się czas obserwacji, wyrażony w latach, który odpowiada wartościom kolumny *Survtime* w tabeli po prawej stronie. Symbol $\circ$ odpowiada wartości kolumny *Status* równej 0, a więc oznacza, że obserwacja była cenzorowana. Symbol $\times$ odpowiada wartości kolumny *Status* równej 1, zatem oznacza wystąpienie zdarzenia (kolumna *Status* to wartość zmiennej losowej c). Pacjent 1 był więc obserwowany przez 7 miesięcy i w tym czasie nie wystąpiło zdarzenie. U pacjenta 2 zdarzenie nastąpiło po 6 miesiącach od początku obserwacji.

Aby określić rozkład przeżycia potrzebujemy następujących definicji funkcji przeżycia oraz funkcji hazardu [13]:

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

$$S(t) = \mathbf{P}(t < T), \quad 0 < t < \infty, \tag{1.2}$$

gdzie: T *- zmienna losowa, oznaczająca czas, w którym nastąpiło zdarzenie.*

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

$$S(0) = 1. \tag{1.3}$$

Funkcję przeżycia często definiuje się także w terminach funkcji hazardu:

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

$$h(t) = lim_{\epsilon \to 0} \frac{\mathbf{P}(t < T < t + \epsilon | T > t)}{\epsilon}. \tag{1.4}$$

Funkcja hazardu jest prawdopodobieństwem tego, 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.

Zdefinowane powyżej dwie funkcje pozwalają na określenie rozkładu przeżycia. Do dalszych analiz przydatnych jest jednak jeszcze kilka definicji [13]:

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

$$F(t) = \mathbf{P}(T \leq t). \tag{1.5}$$

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

$$f(t) = -\frac{d}{dt}S(t) = \frac{d}{dt}F(t). \tag{1.6}$$

Na podstawie powyższych definicji otrzymujemy zależność:

$$h(t) = \frac{f(t)}{S(t)}. \tag{1.7}$$

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* $H : (0, \infty) \to \mathbf{R}$ *dana wzorem:*

$$H(t) = \int_0^t h(u)du. \tag{1.8}$$

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

$$S(t) = \exp\left(-\int_0^t h(u)du\right) = \exp(-H(t)). \tag{1.9}$$

1.2. Modele parametryczne

W analizie przeżycia zakłada się czasami dany rozkład przeżycia, otrzymując model parametryczny. Najprostszym przykładem jest model wykładniczy, w którym zakłada się stały hazard [13]:

$$h(t) = \lambda. \tag{1.10}$$

Wówczas otrzymujemy:

$$H(t) = \int_0^t h(u)du = \int_0^t \lambda du = \lambda t, \tag{1.11}$$

$$S(t) = exp(-H(t)) = exp(-\lambda t), \tag{1.12}$$

$$f(t) = h(t)S(t) = \lambda \exp(-\lambda t). \tag{1.13}$$

Założenie stałego hazardu często nie jest jednak spełnione i szukać należy innych rozkładów estymujących rozkład przeżycia.

Innym często używanym modelem jest model o rozkładzie Weibulla, dla którego funkcja hazardu przyjmuje postać [13]:

$$h(t) = \alpha\lambda(\lambda t)^{\alpha-1}, \quad \alpha, \lambda > 0. \tag{1.14}$$

Dla tego modelu otrzymujemy:

$$H(t) = \int_0^t h(u)du = \alpha\lambda^\alpha \int_0^t u^{\alpha-1}du = \alpha\lambda^\alpha \frac{1}{\alpha} u^\alpha|_0^t = (\lambda t)^\alpha, \tag{1.15}$$

$$S(t) = exp(-H(t)) = exp\left(-(\lambda t)^\alpha\right). \tag{1.16}$$

Rozkład wykładniczy 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 [13]:

$$S(t) = 1 - \Phi\left(\frac{\log(t) - \mu}{\sigma}\right), \tag{1.17}$$

gdzie: Φ - dystrybuanta rozkładu normalnego $\mathcal{N}(0, 1)$. Funkcja hazardu w tym przypadku monotonicznie rośnie od 0 do swojego maksimum, a następnie monotoniczne maleje do 0 przy $t \to \infty$. 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 [13]:

$$f(t) = \frac{\lambda^\beta t^{\beta-1} exp(-\lambda t)}{\Gamma(\beta)}, \quad \lambda, \beta > 0. \tag{1.18}$$

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. Modele nieparametryczne

W wielu przypadkach nie jesteśmy w stanie założyć odpowiedniej rodziny parametrycznej do opisu naszego modelu. Zajmujemy się wtedy modelami nieparametrycznymi. Podstawowym estymatorem funkcji przeżycia używanym w analizie przeżycia jest **estymator Kaplana-Meier'a**, dany wzorem:

$$\hat{S}_{km}(t) = \prod_{t_i \le t} (1 - \hat{q}_i) = \prod_{t_i \le t} \left(1 - \frac{d_i}{n_i}\right), \tag{1.19}$$

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,
N - liczba obserwacji.
Estymator Kaplana-Meier'a jest nierosnącą funkcją schodkową, prawostronnie ciągłą.

Innym sposobem na estymację krzywych przeżycia jest estymator **Fleminga-Harringtona** [7], który opiera się na spostrzeżeniu, że dystrybunata funkcji hazardu 1.1.5 może być przybliżona w następujący sposób:

$$H(t) = \int_0^t h(u)du \approx \sum_{i:t_i \leq t} \tilde{h}_i \Delta_i, \tag{1.20}$$

gdzie:
$t_1, t_2, ..., t_M$ - uporządkowane rosnąco punkty w czasie, w których wystąpiły zdarzenia (bez powtórzeń),
$\tilde{h}_i$ - wartość funkcji hazardu w czasie t_i,
$\Delta_i = t_i - t_{i-1}$ - przedział czasowy, między dwoma kolejnymi zdarzeniami.
Jak zauważono w [7]:

$$h_i \Delta_i \approx \mathbf{P}(t_{i-1} < T < t_i) \approx \frac{d_i}{r_i}, \quad i \in \{1, 2, 3, ..., N\}, \tag{1.21}$$

gdzie:
d_i - liczba zdarzeń, które wystąpiły w czasie t_i,
r_i - liczba jednostek narażonych na ryzyko w czasie t_i (licząc wraz z jednostkiami, u których nastąpiło zdarzenie w czasie t_i).
Oznacza to, że wartość $\tilde{h}_i \Delta_i$ szacuje prawdopodobieństwo wystąpienia zdarzenia w przedziale Δ_i, które może być estymowane poprzez liczbę zdarzeń w czasie t_i podzieloną przez liczbę jednostek narażonych na ryzyko w tym czasie. Estymator dystrybuanty funkcji ryzyka Fleminga-Harringtona wynosi zatem [7]:

$$\hat{H}_{fh}(t) = \sum_{i:t_i \leq t} \frac{d_i}{r_i}. \tag{1.22}$$

Stąd, zgodnie z 1.9, otrzymujemy estymator krzywej przeżycia:

$$\hat{S}_{fh}(t) = \exp(-\hat{H}_{fh}(t)). \tag{1.23}$$

Najczęściej stosowanym estymatorem wariancji dla krzywych przeżycia jest estymator zaproponowany przez Majora Greenwooda [15] w 1926 roku, dany wzorem:

$$\text{var}(\hat{S}(t)) \approx [\hat{S}(t)]^2 \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}. \tag{1.24}$$

Przedział ufności na poziomie istotności α wynosi wówczas:

$$\widehat{CI}(t) = \left[-z_{1-\frac{\alpha}{2}} \sqrt{\text{var}(\hat{S}(t))},\ z_{1-\frac{\alpha}{2}} \sqrt{\text{var}(\hat{S}(t))} \right], \tag{1.25}$$

gdzie: $z_{1-\frac{\alpha}{2}}$ - kwantyl rzędu $1 - \frac{\alpha}{2}$ z rozkładu normalnego.

Takie podejście do estymowania przedziałów ufności, może dać jednak wartości poza przedziałem $[0, 1]$. Aby tego uniknąć, stosuje się transformację `log` funkcji przeżycia. Jak zasugerowano w [14], takie podejście stabilizuje wariancję i dopuszcza niesymetryczne przedziały ufności. Przedział ufności jest wówczas postaci:

$$\text{var}\left(\log \hat{S}(t)\right) = \sum_{i:t_i < t} \frac{d_i}{n_i(n_i - d_i)} \tag{1.26}$$

Jeszcze inną możliwością jest zastosowanie transformacji **log-log** funkcji przeżycia. Dostajemy wówczas [13]:

$$\text{var}\left(log\left[-log\hat{S}(t)\right]\right) \approx \frac{1}{\left[\log \hat{S}(t)\right]^2} \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}. \tag{1.27}$$

W `R` estymacje krzywych przeżycia, można otrzymać za pomocą funkcji `survfit {survival}` [16]. Rodzaj krzywej przeżycia można ustalić za pomocą parametru `type`. Możliwe wartości, to `"kaplan-meier"`, `"fleming-harrington"` oraz `"fh2"`. Wariant `"fh2"` jest modyfikacją przedstawionej powyżej metody Flaminga-Harringtona, w której [16]:

$$\hat{H}_{fh2}(t) = \sum_{i:t_i \leq t} \left(d_i \cdot \sum_{j=1}^{d_i} \frac{1}{r_i - j + 1} \right). \tag{1.28}$$

Funkcja `survfit` dopuszcza także na specyfikację, jaki rodzaj przedziału ufności ma być obliczony. Możliwy wybór, to `"none"` (przedziały ufności nie zostaną obliczone), `"plain"` (1.24), `"log"` (1.26) oraz `"log-log"`(1.27).

1.4. Porównywanie modeli analizy przeżycia

Zagadnieniem, jakim zajmuję się w pracy, jest porównywanie modeli ryzyk konkurujących pośród danych grup obserwacji. Aby wdać się w tego szczegóły należy najpierw zrozumieć ideę porównywania modeli w przypadku analizy przeżycia, którą będziemy później rozszerzać na przypadek wielu możliwych zdarzeń.

W przypadku modeli parametrycznych stosować można 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 funkcje przeżycia. Jako, że porównujemy ze sobą dwie krzywe, test statystyczny ze standardową hipotezą zerową i alternatywną:

$$H_0 : S_1(t) = S_0(t), \tag{1.29}$$
$$H_1 : S_1(t) \neq S_0(t), \tag{1.30}$$

nie jest adekwatny. Dwie krzywe przeżycia, mogą się krzyżować, albo być podobne na jednym odcinku oraz różne na innym odcinku czasu.

Przedstawię kosntrukcję testu porównującego krzywe przeżycia zaproponowanego w [13]. Wprowadzone zostało tutaj rozwiązanie zwane **alternatywą Lehmana**, dla którego hipoteza alternatywna przyjmuje postać:

$$H_1 : S_1(t) = [S_0(t)]^{\psi}. \tag{1.31}$$

Równoważnie, dostajemy test hipotezy zerowej:

$$H_0 : \psi = 1, \tag{1.32}$$

przeciwko hipotezie alternatywnej:

$$H_1 : \psi < 1. \tag{1.33}$$

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 2×2 zawierającą liczbę jednostek u których nastąpiło zdarzenie i u których nie nastąpiło zdarzenie w czasie t_i, dla obydwu grup, tak jak zaprezentowano w tabeli 1.1.

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 zmiennej losowej d_{0i} pod warunkiem n_{0i}, n_i, d_i:

$$\mathbf{P}(d_{0i}|n_{0i}, n_{1i}, d_i) = \frac{\binom{n_{0i}}{d_{0i}}\binom{n_{1i}}{d_{1i}}}{\binom{n_i}{d_i}}, \tag{1.34}$$

gdzie:

$$\binom{n}{d} = \frac{n!}{d!(n-d)!}. \tag{1.35}$$

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

$$e_{0i} = \mathbf{E}d_{0i} = \frac{d_{0i}d_i}{n_i}, \tag{1.36}$$

$$v_{0i} = \mathrm{var}(d_{0i}) = \frac{n_{0i}n_{1i}d_i(n_i - d_i)}{n_i^2(n_i - 1)}. \tag{1.37}$$

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

$$U_0 = \sum_{i=1}^{N}(d_{0i} - e_{0i}), \tag{1.38}$$

$$V_0 = \mathrm{var}(U_0) = \sum_{i=1}^{N} v_{0i}. \tag{1.39}$$

Teraz możemy skonstruować statystykę testową [13]:

$$\frac{U_0^2}{V_0} \sim \chi_1^2. \tag{1.40}$$

Powyższy test nazywany jest **testem log-rank**.

Powyższy test można uogólnić na tak zwany **ważony test log-rang**, taki, że:

$$U_0(w) = \sum_{i=1}^{N} w_i(d_{0i} - e_{0i}), \tag{1.41}$$

$$V_0(w) = \text{var}(U_0(w)) = \sum_{i=1}^{N} w_i^2 v_{0i}. \tag{1.42}$$

Istnieje wiele testów opierających się na powyższej formule, zakładających różne postaci wag. Jednym z nich jest **test Wilcoxona**, dla którego wagą w czasie t_i jest liczba jednostek pod ryzykiem w tym czasie [13]

$$w_i = n_i. \tag{1.43}$$

Test Tarone-Ware'a przypisuje większą wagę do zdarzeń mających miejsce wcześniej, poprzez wykorzystanie pierwiastka z liczby jednostek pod ryzykiem jako wagi [13]:

$$w_i = \sqrt{n_i}. \tag{1.44}$$

Test Flemminga-Harringtona [5] daje największą elastyczność w wyborze statystki testowej, poprzez wybranie parametru ρ:

$$w_i = N \cdot (\hat{S}(t_i))^{\rho}. \tag{1.45}$$

Test Flaminga-Harringtona z parametrem $\rho = 0$ sprowadza się do testu log-rank 1.40.

W środowisku R w pakiecie `survival` [16] porównanie krzywych przeżycia za pomocą testu Flemminga-Harringtona może być wykonane za pomocą funkcji `survdiff`.

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

$$H_0 : h_{0j}(t) = h_{1j}(t), \quad \text{dla } j \in \{1, 2, ..., n_G\}. \tag{1.46}$$

Dla każdej wartości zmiennej $G = g$ obliczamy statystyki U_{0g} oraz V_{0g} a następnie wyznaczamy statystykę testową [12]:

$$X^2 = \frac{(\sum_{n=1}^{n_g} U_{0g_n})^2}{\sum_{n=1}^{n_g} V_{0g_n}^2} \sim \chi^2_{n_g - 1}. \tag{1.47}$$

Statystyka testowa w powyższym teście różni się od tej z testu log-rank 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` zaimplemetnowana 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**:

$$h_1(t) = \Psi h_0(t), \tag{1.48}$$

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 [12]:

$$h_1(t, z_1, z_2, ..., z_m) = h_0(t)e^{\beta_1 z_1 + \beta_2 z_2 + ... + \beta_m z_m}, \tag{1.49}$$

gdzie:
$h_0(t)$ - hazard bazowy,
$z_1, z_2, ..., z_m$ - zmienne objaśniane, niezależne od czasu t,
$\beta_1, \beta_2, ..., \beta_m$ - parametry.
Model Coxa nazywany jest czasem **modelem regresji Coxa**, ponieważ można go w łatwy sposób sprowadzić do postaci liniowej:

$$\log \frac{h_1(t, z_1, z_2, ..., z_m)}{h_0(t)} = \beta_1 z_1 + \beta_2 z_2 + ... + \beta_m z_m. \tag{1.50}$$

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 **semiparametryczny**), co czyni go adekwatnym do modelowania w różnych sytuacjach. Jeżeli nie jesteśmy pewni, co do 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 $\beta_1, ..., \beta_m$. Dzięki tym własnościom jest on najczęściej używanym modelem w analizie przeżycia.

Estymację współczynników β_i można wykonać przy pomocy metody największej wiarygodności. Jako, że nie potrzebujemy wiedzy na temat postaci hazardu bazowego, korzystamy tutaj z częściowej wiarygodności [13]:

$$\alpha(\beta) = \prod_{i:c_i=1} \frac{e^{Z_i\beta_i}}{\sum_{j:t_j>t_i} e^{Z_1\beta_1 + ... + Z_m\beta_m}}. \tag{1.51}$$

gdzie $\beta = (\beta_1, \beta_2, ..., \beta_m)$.

Przedstawię 3 testy na istotność współczynników w modelu Coxa. Pierwszym z nich jest **test ilorazu wiarygodności**, w którym statystyke testowa wynosi [13]:

$$LRT = 2\left[\log \alpha(\hat{\beta}) - \log \alpha_0\right] \sim \chi^2_1, \tag{1.52}$$

gdzie:
α_0 - wiarygodnością modelu zerowego (ze wszystkimi parametrami równymi 0),
$\hat{\beta}$ jest estymatorem β.

Test Walda przyjmuje postać [13]:

$$W = \frac{\hat{\beta}}{se(\hat{\beta})} \sim \mathcal{N}(0,1), \tag{1.53}$$

gdzie $se(\hat{\beta})$ jest błędem standardowym estymatora $\hat{\beta}$.

Test mnożników Lagrange'a (ang. *the score test*) przyjmuje postać [13]:

$$\frac{\frac{d}{dt}\log\alpha(\beta_0)}{\sqrt{\text{var}(\log\alpha(\beta_0))}}, \tag{1.54}$$

gdzie β_0 jest wartością β z hipotezy zerowej. Może on zostać wykonany bez znalezienia estymatora największej wiarygodności $\hat{\beta}$.

W `R` estymację modeli Coxa oraz wyniki powyższych testów można otrzymać za pomocą funkcji `coxph` {`survival`} [16].

1.6. Modele ryzyk konkurujących

Modelami, którymi zajmuję się w tej pracy są **modele ryzyk konkurujących** (ang. *competing risks models*). Jest to jedno z dwóch, obok modeli wielostanowych (ang. *multistate models*), uogólnień analizy przeżycia, dopuszczających występowanie więcej niż jednego ryzyka. W przypadku modeli ryzyk konkurujących mamy do czynienia z więcej niż jednym możliwym zdarzeniem i obserwujemy czas do wystąpienia pierwszego z nich, w odróżnieniu od 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 komunikacji miejskiej lub roweru jako środka transportu do pracy.

Formalnie, określamy nasz **zbiór ryzyk** jako $\{1,2,3,...,J\}$. Każde z nich jest określone jednoznacznie (mogą to być np. różne przyczyny śmierci).

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 przykładach w następnych rozdziałach.

1.7. Funkcje skumulowanych częstości

Pierwszym prezentowanym tutaj podejściem stosowanym do modelowania ryzyk konkurujących jest zastosowanie tak zwanych **funkcji skumulowanych częstości** (ang. *cumulative incidende function*, także *subdistribution function*) dla każdego z ryzyk.

Definicja 1.7.1 ***Funkcja skumulowanych częstości*** *- funkcja $F_j : [0,\infty) \to [0,1]$, dana wzorem:*

$$F_j(t) = \mathbf{P}(T \leq t, \delta = j) = \int_0^t h_j(u)S(u)du, \quad j \in \{1,2,3,...,J\}, \tag{1.55}$$

gdzie δ oznacza dany typ zdarzenia.

Funkcja ta posiada pewne analogie do dystrybuanty funkcji ryzyka, jednak jej granica przy $t \to \infty$ jest równa prawdopodobieństwu wystąpienia danego zdarzenia, a nie zbiega do 1. Dokładniej:

$$lim_{t\to\infty} F_j(t) = \mathbf{P}(\delta = j), \quad j \in \{1, 2, 3, ..., J\}. \tag{1.56}$$

Analogicznie definiujemy funkcję hazardu dla danego ryzyka j:

$$h_j(t) = lim_{\epsilon\to 0} \frac{\mathbf{P}(t < T < t + \epsilon, \delta = j | T > t)}{\epsilon}, \quad j \in \{1, 2, 3, ..., J\}. \tag{1.57}$$

Dodając do siebie funkcje hazardu dla poszczególnych ryzyk, dostajemy ogólną funkcję hazardu:

$$h(t) = \sum_{j=1}^{J} h_j(t). \tag{1.58}$$

Wzór ten ma ważną interpretację, oznaczającą, iż ryzyko wystąpienia jednego ze zdarzeń w danym punkcie czasowym jest sumą ryzyk wystąpienia poszczególnych zdarzeń w tym czasie.

Estymację funkcji hazardu można przeprowadzić w sposób analogiczny do przypadku analizy przeżycia. Dla $\{t_1, t_2, ..., t_N\}$ - uporządkowanych czasów wystąpienia zdarzeń, funkcja hazardu dla danego ryzyka wynosi:

$$\hat{h}_j(t_i) = \frac{d_{ij}}{n_i}, \quad j \in \{1, 2, 3, ..., J\}, \tag{1.59}$$

gdzie:
d_{ij} - liczba zdarzeń typu j które wystąpiły w czasie t_i,
n_i to liczba jednostek narażonych na ryzyko w czasie t_i.
Otrzymujemy wówczas:

$$\hat{h}(t_i) = \frac{\sum_{j=1}^{J} d_{ij}}{n_i} = \sum_{j=1}^{J} \hat{h}_j(t_i). \tag{1.60}$$

Estymator funkcji skumulowanych częstości wynosi wtedy:

$$\hat{F}_j(t) = \sum_{t_i \le t} \hat{S}(t_{i-1}) \hat{h}_j(t_i). \tag{1.61}$$

Test na porównywanie krzywych skumulowanych częstości pomiędzy grupami, tak zwany **test dla K prób** (ang. *K-sample test*), został po raz pierwszy zaproponowany przez Roberta J. Greya w 1988 roku [9] i stanowi on analogię do testu logrank 1.40. Przedstawię konstrukcję tego testu na podstawie [9]. Zakładamy, każda jednostka należy do jednej z grup $\{1, 2, 3, ..., K\}$ i dane są prawostronnie cenzorowane. Przyjmijmy następującą notację:

- T_{ik}^0 - czas zdarzenia dla $i - tej$ jednostki z grupy k, $i \in \{1, 2, ..., n_k\}$,
- $n = \sum_{k=1}^{K} n_k$ - liczba wszystkich obserwacji,
- $\delta_{ik}^0 \in \{1, 2, 3, ...J\}$ - typ zdarzenia dla $i - tej$ jednostki z grupy k, $i \in \{1, 2, ..., n_k\}$,
- $F_{jk}(t) = \mathbf{P}(T_{ik}^0 \le t, \delta_{ik}^0 = j)$ - funkcja skumulowanych częstości dla ryzyka j w grupie k,
- $f_{jk}(t) = \frac{d}{dt} F_{jk}(t)$ - funkcja gęstości dla skumulowanej częstości.

Zakładamy, że pary $(T_{ik}^0, \lambda_{ik}^0)$ dla danego ryzyka j są niezależne, o jednakowym rozkładzie. Nie zakładamy jednak niezależności ryzyk. Dla ułatwienia notacji przyjmujemy, że interesującym nas typem zdarzenia jest $j = 1$. Hipotezą zerową naszego testu jest wówczas:

$$H_0 : F_{1k} = F_1^0, \quad k \in \{1, 2, 3, ...K\}, \tag{1.62}$$

gdzie: F_1^0 jest niesprecyzowaną funkcją skumulowanych częstości. W terminach zdefiniowanych powyżej, funkcja przeżycia w grupie k przyjmuje postać:

$$S_k(t) = \mathbf{P}(T_{ik}^0 > t) = 1 - \sum_{j=1}^{J} F_{jk}(t). \tag{1.63}$$

Funkcja hazardu dla zdarzenia typu j w grupie k wynosi wówczas:

$$\lambda_{jk}(t) = \frac{f_{jk}(t)}{S_k(t)}. \tag{1.64}$$

Dla ułatwienia notacji przedstawmy konstrukcję testu dla $J = 2$ ryzyk. Nie narzucamy tym samym żadnych ograniczeń, gdyż w przypadku więcej niż dwóch ryzyk, możemy testować różnice między jednym ryzykiem, a drugim będącym *wszystkimi innymi ryzykami*. Główną ideą tego testu jest porównanie **ważonych hazardów subdystrybucyjnych**:

$$\gamma_{jk}(t) = \frac{f_{jk}(t)}{1 - F_{jk}(t)} = \frac{f_{jk}(t)}{G_{jk}(t)}, \tag{1.65}$$

gdzie: $G_{jk}(t) = 1 - F_{jk}(t)$. Analogicznie jak w przypadku podstawowych metod analizy przeżycia, możemy zdefiniować dystrybuantę hazardu subdystrybucyjnego:

$$\Gamma_{jk}(t) = \int_0^t \gamma_{jk}(u)du. \tag{1.66}$$

Przez U_{ik} oznaczmy czas cenzorowania dla $i-tej$ jednostki w grupie k. Tak jak zauważyliśmy w rozdziale 1.1, zakładamy, że U_{ik} są niezależne od $(T_{ik}^0, \delta_{ik}^0)$. Wartości obserwowane, to:

$$T_{ik} = \min(T_{ik}^0, U_{ik}), \tag{1.67}$$

$$\delta_{ik} = \delta_{ik}^0 \mathbf{I}(T_{ik} \leq U_{ik}). \tag{1.68}$$

Konstrukcja testu opiera się na teorii procesów liczących (ang. *counting processes*) zaprezentowanej w [2]. Zdefiniujmy zliczenia zdarzeń $j - tego$ rodzaju w $k - tej$ grupie, które wystąpiły przed czasem t jako:

$$N_{jk}(t) = \sum_{i=1}^{n_k} \mathbf{I}(T_{ik} \leq t, \delta_{ik} = j) \tag{1.69}$$

oraz zliczenia jednostek narażonych na ryzyko w $k - tej$ grupie, po czasie t:

$$Y_k(t) = \sum_{i=1}^{n_k} \mathbf{I}(T_{ik} \geq t). \tag{1.70}$$

Wówczas możemy skonstruować następujący estymator funkcji skumulowanych częstości:

$$\hat{F}_{jk}(t) = \int_0^t \hat{S_{km}}(u-) Y_k^{-1}(u) dN_{jk}(u), \tag{1.71}$$

gdzie: $\hat{S_{km}}(t)$ jest estymatorem Kaplana-Meiera zdefiniowanym tak jak w 1.19, $\hat{S}_k(t-) = lim_{s \to t^-} \hat{S}_k(s)$ oraz przyjmujemy $\hat{S}_k(t-) = 0$ dla $Y_k(t) = 0$ (z przyjętą konwencją $\frac{0}{0} = 0$).

Jak pokazano w [1], przy założeniu niezależności ryzyk, estymator ten jest silnie zgodny i słabo zbieżny, co więcej jest on estymatorem największej wiarygodności dla modeli nieparametrycznych [10].

Zauważmy, że na podstawie danych nie jesteśmy w stanie obliczyć rozkładu F_1^0 z hipotezy zerowej, gdyż nie zakłada ona, że S_k ani λ_{1k} mają być równe w różnych grupach. Stąd definiujemy **zbiór ryzyka**, jako:

$$R_k(t) = \frac{\mathbf{I}(\tau_k \geq t) Y_k(t) \hat{G}_{1k}(t-)}{\hat{S}_k(t-)}, \tag{1.72}$$

gdzie: $\tau_k, k \in \{1, 2, 3, ..., K\}$ to ustalone czasy, które spełniają, przy założeniu hipotezy zerowej:

$$\Pi_k^0(t) = \alpha_k \mathbf{P}(T_{ik} \geq t) > 0, \text{ dla } 0 < \alpha_k < \frac{n_k}{n}. \tag{1.73}$$

Przyjmując, że $R_k(t) = 0$ dla $\tau_k < t$, otrzymujemy:

$$\hat{\Gamma}_{1k}(t) = \int_0^t [\hat{G}_{1k}(u-)]^{-1} d\hat{F}_{1k}(u) = \int_0^t [R_k(u)]^{-1} dN_{1k}(u), \quad \text{dla} \quad t \leq \tau_k, \tag{1.74}$$

przy czym ostatnia równość wynika z 1.71. Za estymator Γ_1^0 przyjmujemy wówczas:

$$\hat{\Gamma}_1^0(t) = \int_0^t [R_\cdot(u)]^{-1} dN_{1\cdot}(u), \tag{1.75}$$

gdzie indeks $'\cdot'$ oznacza sumowanie po wszystkich możliwych wartościach k. Estymator ten jest zgodny przy założeniu hipotezy zerowej, ponieważ wszystkie estymatory $\hat{F}_{1k}$ zgodnie estymują F_1^0 oraz:

$$\hat{\Gamma}_1^0(t) = \sum_{k=1}^K \int_0^t \Big[\frac{R_k(u)}{R_\cdot(u)}\Big] \hat{G}_{1k}^{-1}(u-) d\hat{F}_{1k}(u). \tag{1.76}$$

Ostatecznie, jako statystykę testową przyjmujemy:

$$z_k = \int_0^{\tau_k} K_k(t) \left[d\hat{\Gamma}_{1k} - d\hat{\Gamma_1^0} \right], \tag{1.77}$$

gdzie $K_k(t)$ jest wybraną funkcją wag, zazwyczaj postaci $K_k(t) = L(t) R_k(t)$, dla pewnej funkcji $L(T)$.

Jak zostało udowodnione w [9], zakładając, że $K_k(t)$ jest procesem na $[0, \tau_k]$, zbiegającym jednostajnie, według prawdopodobieństwa do $K_k^0(t)$, mamy słabą zbieżność statystyki testowej:

$$n^{-\frac{1}{2}} Z \to N_k(\mu, \Sigma). \tag{1.78}$$

gdzie $Z = (z_1, z_2, ..., z_K)'$, dla pewnych parametrów μ i Σ.

W przypadku tylko $K = 2$ grup, przeprowadzenie testu sprowadza się do obliczenia:

$$\int_0^\tau K(t) \left([1 - \hat{F_{11}}(t-)]^{-1} d\hat{F_{11}}(t) - [1 - F_{12}\hat{(t-)}]^{-1} d\hat{F_{12}}(t) \right), \tag{1.79}$$

gdzie ponownie $K(t)$ jest wybraną funkcją wag.

W środowisku `R`, funkcje skumulowanych częstości oraz test dla K prób można obliczyć za pomocą funkcji `cuminc` z pakietu `cmprsk` [8].

1.8. Model Coxa dla ryzyk konkurujących

Innym podejściem pozwalającym na modelowanie ryzyk konkurujących jest zastosowanie uogólnionego modelu Coxa, zaprezentowanego po raz pierwszy przez J. Fine'a oraz R. Grey'a w 1999 roku [6]. Możemy tutaj przyjąć uproszczoną notację:

- T - czas zdarzenia,
- C - czas cenzorowania,
- $\delta \in \{1, 2, 3, ..., J\}$ - typ zdarzenia ,
- Z - wektor współczynników długości m, $z_i, i \in \{1, 2, 3, ..., m\}$ ograniczone, niezależne od czasu.

Dla danych prawostronnie cenzorowanych obserwujemy: $X = \min(T, C)$, $\Delta = \mathbf{I}(T \leq C)$ oraz Z. Zakładamy, że $\{X_i, \Delta_i, \Delta_i\delta_i, Z_i\}$ są niezależne, o jednakowym rozkładzie dla $i \in \{1, 2, 3, ..., N\}$, gdzie N jest liczbą obserwacji. Ponownie zakładamy, że interesuje nas model dla zdarzenia typu 1. Funkcja skumulowanych częstości przyjmuje wówczas postać:

$$F_1(t; Z) = \mathbf{P}(T \leq t, \delta = 1 | Z). \tag{1.80}$$

Używać będziemy klasy semiparametrycznych transformacji modelu, to znaczy, rozważamy pewną funkcję rosnącą g, taką, że:

$$g(F_1(t; z)) = h_0(t) + Z^{\mathbf{T}} \beta_0, \tag{1.81}$$

gdzie:
$h_0(t)$ - niesprecyzowana, monotonicznie rosnącą funkcją,
β_0 - wektor parametrów długości p.
Najczęstszym wyborem funkcji g jest:

$$g(u) = \log(-\log(1 - u)). \tag{1.82}$$

Podobnie jak w przypadku testu dla K prób definiujemy hazard subdystrybucyjny:

$$\begin{aligned} \lambda_1(t; Z) &= lim_{\epsilon \to 0} \frac{\mathbf{P}(t \leq T \leq t + \epsilon, \delta = 1 | T \geq t \cup (T \leq t \cap \delta \neq 1), Z)}{\epsilon} \\ &= \frac{\frac{dF_1(t;Z)}{dt}}{1 - F_1(t; Z)} = -\frac{d \log(1 - F_1(t; Z))}{dt}. \end{aligned} \tag{1.83}$$

Przy założeniu proporcjonalenego hazardu mamy:

$$\lambda_1(t;Z) = \lambda_{10}(t;Z)\exp(Z^{\mathbf{T}}\beta_0), \tag{1.84}$$

gdzie $\lambda_{10}(t)$ jest niesprecyzowaną, nieujemną funkcją czasu. Użycie transformacji $g(u) = \log(-log(u))$ daje nam hazard bazowy postaci:

$$h_0(t) = \log\left(\int_0^t \lambda_{10}(s)ds\right), \tag{1.85}$$

dzięki czemu zarówno hazard bazowy jak i współczynniki regresji mają prostą interpretację niezależną od struktury hazardu subdystrybucyjnego.

Testowanie różnic pomiędzy grupami w modelach Coxa w przypadku występowania ryzyk konturujących można wykonać za pomocą **modyfikowanego testu ilorazy wiarygodności**. Aby obliczyć częściową wiarygodność dla naszego modelu definiujemy zbiór ryzyka dla $i-tej$ jednostki jako:

$$R_i = \{k : (\min(C_k, T_k) \geq T_i) \cup (T_k \leq T_i \cap \delta_k \neq 1 \cap C_k \geq T_i)\}. \tag{1.86}$$

Częściowa wiarygodność wynosi wówczas [6]:

$$\alpha_{cr}(\beta) = \Pi_{i=1}^{n}\left[\frac{\exp(Z_i^{\mathbf{T}}(T_i)\beta)}{\sum_{k\in R_i}\exp(Z_j^{\mathbf{T}}(T_i)\beta)}\right]. \tag{1.87}$$

Statystyka testowa, wynosi wówczas:

$$LRT_{cr} = 2\left[\log\alpha_{cr}(\hat{\beta}) - \log\alpha_{cr}^0\right] \sim \chi_1^2, \tag{1.88}$$

gdzie: α_{cr}^0 to częściowa wiarygodność dla modelu zerowego.

Estymację modelu Coxa dla ryzyk konkurujących w `R` można uzyskać za pomocą funkcji `crr` z pakietu `cmprsk` [8]. Funkcja ta oblicza także częściową wiarygodność dla danych prawostronnie cenzorowanych dla wyestymowanego modelu oraz dla modelu zerowego (z wszystkimi współczynnikami β_0 równymi 0), co umożliwia obliczenie modyfikowanego testu ilorazu wiarygodności.

Rozdział 2

Biblioteka 'cr17'

2.1. Wprowadzenie

Biblioteka `cr17` stanowi narzędzie do analizy i wizualizacji modeli ryzyk konkurujących. Głównym punktem zainteresowania jest badanie różnic między modelami dla poszczególnych zdarzeń pośród określonych grup w populacji. Dostępne funkcje opierają się na tych zaimplementowanych w pakietach `survival` [16] oraz `cmprsk` [8], posiadają one jednak liczne udogodnienia i są przystosowane na przypadek ryzyk konkurujących. Wizualizacje są wykonywane przy użyciu pakietu `ggplot2`.

Podczas pracy nad biblioteką dążyłam do stworzenia narzędzia przystępnego dla każdego użytkownika. Klarowność pakietu wynika z jego następujących cech:

- brak konieczności wywoływania tej samej funkcji wiele razy, dla każdego ze zdarzeń osobno, jak w przypadku innych pakietów zajmujących się modelami ryzyk konkurujących,
- brak restrykcji co do typu wektorów zawierających dane o rodzaju zdarzenia i grupie, do której należy obserwacja (w niektórych pakietach spotykamy np. ograniczenie na numeryczną zmienną określającą rodzaj ryzyka, co utrudnia prace poprzez konieczność zakodowania zmiennej typu `character` lub `factor` na zmienną numeryczną. Do stworzenia legend przy wykresach czy starannego raportu, należy ponownie powrócić do pierwotnych nazw),
- funkcja `summarizeCR` pozwalająca na uzyskanie kompleksowego raportu poprzez wywołanie tylko jednej linijki kodu.

Do pakietu dołączone zostały dane `LUAD` pochodzące z badania *The Cancer Genome Atlas* [11], dotyczące zgonów oraz nawrotów chorób wśród pacjentów z nowotworami płuc. Na podstawie tych danych przedstawię funkcjonalność pakietu.

Pakiet składa się z 12 funkcji, które estymują poszczególne modele, wykonują testy diagnostyczne oraz tworzą tabele i wykresy dla dwóch podejść - analizy przeżycia, w której zdarzenia innego rodzaju traktujemy jako cenzorowanie oraz dla modeli ryzyk konkurujących. Dodatkowo, zaimplementowana została funkcja `summarizeCR`, generująca sumaryczny raport bez konieczności wywoływania poszczególnych funkcji.

Aby przejrzyście przedstawić działanie pakietu, przyjrzyjmy się najpierw danym `LUAD`. Zawierają one informacje o czasie obserwacji, rodzaju zdarzenia oraz płci, którą traktować będziemy jako zmienną grupującą, pośród 522 pacjentów. Dokładniej, dane zawierają 3 kolumny:

- **event**, określająca, czy dana jest cenzorowana (*alive*), czy wystąpiło jedno z dwóch konkurujących zdarzeń - zgon (*death*) lub pojawienie się nowego nowotworu (*new_tumor*),
- **time**, liczba dni od początku obserwacji do wystąpienia zdarzenia, bądź końca obserwowania, w przypadku cenzorowania,
- **gender**, płeć pacjenta.

W tabeli 2.1 przedstawione zostały liczebności zdarzeń w poszczególnych grupach.

Tablica 2.1: Tabela liczebności dla danych `LUAD`.

	Male	*Female*	Razem
alive	182	207	389
death	46	56	102
new_tumor	14	17	31
Razem	242	280	522

W bibliotece `cr17` rozważone zostały 4 podejścia modelowania ryzyk konkurujących. Pierwsze dwa, to estymacja krzywych przeżycia oraz modele Coxa dla poszczególnych ryzyk, podczas gdy zdarzenia innego typu traktowane są jako cenzorowane. Następne dwa to estymacja funkcji skumulowanych częstości oraz modele Coxa dla ryzyk konkurujących. W następnych podrozdziałach opisane zostaną zaimplementowane funkcje.

2.2. Estymacja modeli analizy przeżycia - funkcja fitSurvival

Funkcja `fitSurvival` jest adaptacją funkcji `survfit` z pakietu `survival`, dostosowaną do przypadku więcej niż jednego ryzyka. Dopasowuje ona krzywe przeżycia, opisane w rozdziale 1.3, dla każdego z ryzyk i każdej z grup osobno, traktując obserwacje, u których wystąpiło zdarzenie innego rodzaju, jako cenzorowane. Argumentami tej funkcji są:

- **time**, wektor zawierający punkty w czasie, w ktorych wystąpiło zdarzenie, bądź nastąpił koniec obserwacji. Kolumna ta powinna być typu `numeric`,
- **risk**, wektor określający typ zdarzenia, może być typu `numeric`, `character` lub `factor`,
- **group**, wektor określający grupę, do której należy dana obserwacja, może być typu `numeric`, `character` lub `factor`,
- **cens**, wartość oznaczająca obserwacje cenzorowane w kolumnie **risk** (domyślnie *NULL*, przyjęta zostanie pierwsza wartość z wektora **risk**),
- **type**, rodzaj krzywej przeżycia, jaka ma być obliczona. Możliwe wartości to: `"kaplan-meier"` (wartość domyślna), `"fleming-harrington"` oraz `"fh2"` (patrz: rozdział 1.3),
- **conf.int**, poziom ufności (domyślnie 0.95),

– **conf.type**, rodzaj przedziału ufności, jaki ma być obliczony. Możliwe wartości to: `none` (przedziały ufności nie zostaną obliczone), `plain`, `log` (wartość doymślna), `log-log` (patrz: rozdział 1.3).

Argumenty **time**, **risk**, **group**, **cens** są argumentami większości funkcji i nie będą ponownie opisywane przy omawianiu kolejnych funkcji.

Wartością funkcji `fitSurvival` jest lista, której elementami są objekty klasy `survfit.summary {survival}` dla poszczególnych ryzyk. Każdy taki element jest listą, zawierającą w szczególności następujące informacje:

– **time**, czasy kolejnych zdarzeń danego ryzyka,

– **n.risk**, liczba jednostek narażonych na ryzyko w czasie **time**,

– **n.event**, liczba zdarzeń w czasie **time**,

– **surv**, wartość estymowana krzywej przeżycia w czasie **time**,

– **strata**, grupa, do której należy dana obserwacja,

– **std.err**, błąd standardowy estymacji krzywej przeżycia w czasie **time**,

– **lower**, dolne organiczenie przedziału ufności dla krzywej przeżycia w czasie **time**,

– **upper**, górne ograniczenie przedziału ufności dla krzywej przeżycia w czasie **time**.

Po wywołaniu następującego kodu:

```
fitS <- fitSurvival(time = LUAD$time,
                    risk = LUAD$event,
                    group = LUAD$gender,
                    cens = "alive",
                    type = "kaplan-meier",
                    conf.int = 0.95,
                    conf.type = "log")
```

otrzymujemy dwuelementową listę. Pierwszym elementem tej listy jest objekt typu `survfit.summary` dla ryzyka *death*, drugim dla ryzyka *new_tumor*.

2.3. Rysowanie krzywych przeżycia - funkcja plotSurvival

Wynik funkcji `fitSurvival` dostarcza nam wystarczających informacji do narysowania krzywych przeżycia dla poszczególnych ryzyk pośród danych grup. Służy do tego funkcja `plotSurvival`, której argumentami są:

– **fit**, objekt powstały po wywołaniu funkcji `fitSurvival`,

– **target**, punkt w czasie, dla którego narysowane zostaną przedziały ufności na wykresie (domyślnie *null*, przedziały ufności nie zostaną narysowane).

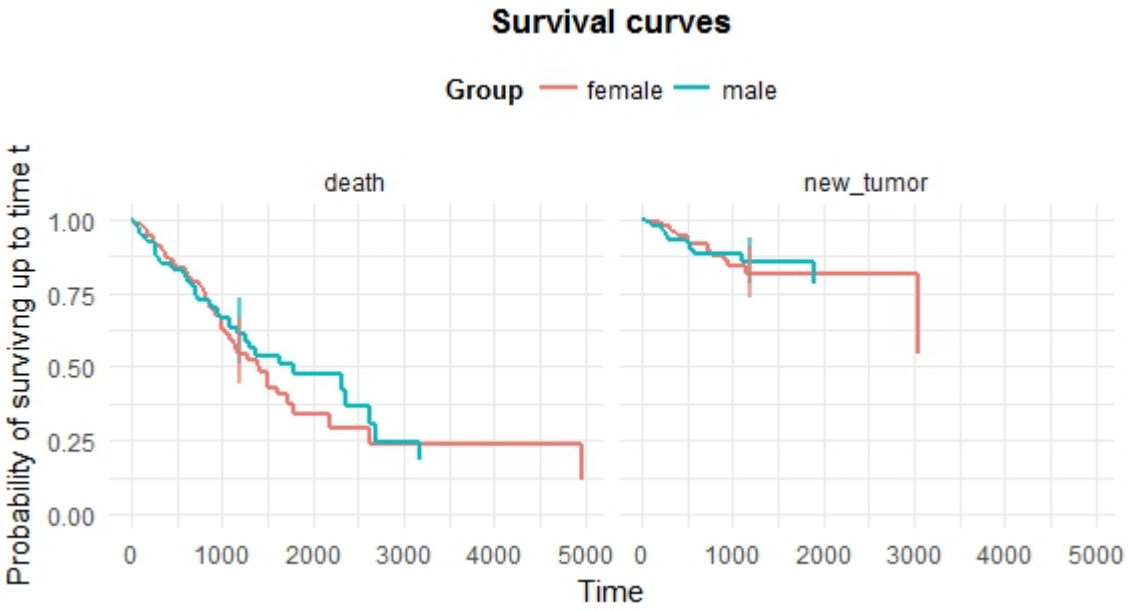
– **ggtheme**, argument funkcji ggplot, wybór tematu wykresu (domyślnie: *theme_minimal*()).

– **titleSurv**, tytuł wykresu (domyślnie: *Survival curves*).

– **xtitle**, nazwa osi pionowej (domyślnie: *Time*).

– **ytitleSurv**, nazwa osi poziomej (domyślnie: *Probability of survivng up to time t*).

– **legendtitle**, tytuł legendy (domyślnie: *Group*).



Rysunek 2.1: Wynik funkcji `plotSurvival` na danych LUAD. Wykres przedstawia krzywe przeżycia wśród pacjentów u których pierwszym zdarzeniem był zgon (*death*) oraz u których pierwszym zdarzeniem było wystąpienie nowego nowotworu (*new_tumor*). Krzywe przeżycia są rysowane osobno dla kobiet (*female*) oraz dla mężczyzn (*male*).

Wynikiem tej funkcji jest wykres przedstawiający krzywe przeżycia. Po wywołaniu:

```
plotSurvival(fit = fitS,
             target = 1200,
             ggtheme = theme_minimal(),
             titleSurv = "Survival curves",
             xtitle = "Time",
             ytitleSurv = "Probability of survivng up to time t",
             legendtitle = "Group")
```

otrzymujemy wykres jak na rysunku 2.1.

2.4. Testowanie modeli analizy przeżycia - funkcja testSurvival

Wyniki testu Fleminga-Hrringtona (patrz: rozdział 1.4), badającego istotność różnić w krzywych przeżycia w grupach, otrzymujemy za pomocą funkcji `testSurvival`. Poza standardowymi argumentami przyjmuje ona także parametr **rho**, zdefiniowany w 1.45. Domyślna wartość tego parametru wynosi 0 (otrzymujemy wówczas wyniki testu logrank). Wynikiem tej funkcji jest tabelka `data.frame`, zawierająca p-wartości dla testu Fleminga-Harringtona dla poszczególnych ryzyk. Dla danych LUAD mamy:

```
testSurvival(time = LUAD$time,
             risk = LUAD$event,
             group = LUAD$gender,
             cens = "alive",
             rho = 0)
```

Wynik z funkcji `testSurvival` na danych LUAD został zaprezentowany na rysunku 2.2.

	death	new_tumor
Fleming-Harrington test	0.76	0.99

Rysunek 2.2: Wynik funkcji `testSurvival` na danych LUAD. W tabeli znajdują się p-wartości testu Fleminga-Harringtona (patrz: rozdział 1.4), badającego istotność występowania różnic w krzywych przeżycia wśród kobiet i mężczyzn dla obydwu ryzyk.

2.5. Estymacja modeli Coxa - funkcja fitCox

Funkcja `fitCox` dopasowuje model Coxa (patrz: rozdział 1.5), na podstawie funkcji `coxph` z pakietu `survival`, dla poszczególnych ryzyk, traktując inne zdarzenia jako cenzorowane. Argumentami są **time**, **risk**, **group**, **cens**, **conf.int** zdefiniowane powyżej. Wynikiem jest lista, której elemantami są objekty klasy `coxph.summary`. Każdy z tych elementów, zawiera następujące informacje:

- dopasowane współczynniki *beta* (patrz: rozdział 1.5),
- przedziały ufności dla tych współczynników,
- statystykę testową i p-wartość dla testu ilorazu wiarygodności 1.52,
- statystykę testową i p-wartość dla testu Walda 1.53,
- statystykę testową i p-wartość dla testu mnożników Lagrange'a 1.54.

Po wywołaniu:

```
fitC <- fitCox(time = LUAD$time,
               risk = LUAD$event,
               group = LUAD$gender,
               cens = "alive",
               conf.int = 0.95)
```

otrzymujemy dwuelementową listę, której elementami są obiekty klasy `coxph.summary` dla poszczególnych ryzyk.

2.6. Testowanie modeli Coxa - funkcja testCox

Po dopasowania modelu Coxa dla poszczególnych ryzyk, za pomocą funkcji `testCox` otrzymujemy p-wartości dla trzech testów badających różnice pomiędzy grupami: testu ilorazu wiarygodności 1.52, testu Walda 1.53 oraz testu logrank 1.54. Argumentem tej funkcji jest `fitCox` - wynik funkcji `fitCox`. Po wywyołaniu

```
testCox(fitCox = fitC)
```

dostajemy tabelkę z p-wartościami dla wyżej wymienionych testów dla obydwu ryzyk. Wynik funkcji `testCox` na danych `LUAD` został zaprezentowany na rysunku 2.3.

	death	new_tumor
Likelihood ratio test	0.76	0.99
Wald test	0.76	0.99
Logrank test	0.76	0.99

Rysunek 2.3: Wynik funkcji `testCox` na danych LUAD. W tabeli znajdują się p-wartości testów badających istotność różnic pomiędzy modelami Coxa dla kobiet i dla mężczyzn. Wykonane testy to: **LRT** - test ilorazu wiarygodności dla modeli Coxa 1.52, **Wald Test** - test Walda 1.53, **Logrank Test** - test logrank dla modeli Coxa 1.54.

2.7. Estymacja modeli ryzyk konkurujących - funkcja fitCuminc

Funkcja `fitCuminc` estymuje funkcje skumulowanych częstości w poszczególnych grupach, za pomocą funkcji `cuminc` z pakietu `cmprsk`. Argumentami tej funkcji są `time`, `risk`, `group`, `cens`. Otrzymaną wartością jest lista, której elementami są oszacowania krzywych skumulowanych gęstości dla poszczególnych grup i ryzyk. Każdy z tych elementów, zawiera następujące informacje:

- **time**, punkty w czasie, w których wystąpiły zdarzenia danego typu w danej grupie,
- **est**, estymowana wartość krzywej skumulowanych gęstości w czasie **time**,
- **var**, wariancja estymowanej wartości funkcji skumulowanych częstości.
- **group**, grupa, dla której estymowana jest funkcja skumulowanych częstości,
- **risk**, typ zdarzenia, dla którego estymowana jest funkcja skumulowanych częstości.

Dodatkowym elementem jest tabelka `data.frame` zawierająca wyniki testu dla K-prób (patrz: rozdział 1.7). Po wywołaniu:

```
fitC <- fitCuminc(time = LUAD$time,
                  risk = LUAD$event,
                  group = LUAD$gender,
                  cens = "alive")
```

otrzymujemy pięcioelementową listę. Pierwsze cztery elementy to oszacowania krzywych skumulowanych gęstości dla każdego ryzyka w każdej grupie. Piątym elementem, jest tabelka zawierająca dane na temat statystyki testowej oraz p-wartości testu dla K-prób dla obydwu ryzyk.

2.8. Rysowanie krzywych skumulowanych częstości - funkcja plotCuminc

Na podstawie funkcji `fitCuminc` można otrzymać wykres skumulowanych wartości, za pomocą funkcji `plotCuminc`. Jej argumentami są

- **ci**, objekt powstały po wywołaniu funkcji `fitCuminc`,
- **cens**, wartość wektora. zawierającego typ zdarzenia, oznaczająca obserwację cenzorowaną,
- **target**, punkt w czasie, dla którego narysowane zostaną przedziały ufności na wykresie.
- **ggtheme**, argument funkcji ggplot, wybór tematu wykresu (domyślnie: *theme_minimal*()).
- **titleCuminc**, tytuł wykresu (domyślnie: *Cumulative incidence functions*).
- **xtitle**, nazwa osi pionowej (domyślnie: *Time*).
- **ytitleCuminc**, nazwa osi poziomej (domyślnie: *Cumulative incidences*).
- **legendtitle**, tytuł legendy (domyślnie: *Group*).

Po wywołaniu:

```
plotCuminc(ci = fitC ,
           cens = "alive",
           target = 1200,
           ggtheme = theme_minimal(),
           titleCuminc = "Cumilative incidence function",
           xtitle = "Time",
           ytitleCuminc = "Cumulative incidences",
           legendtitle = "Group")
```

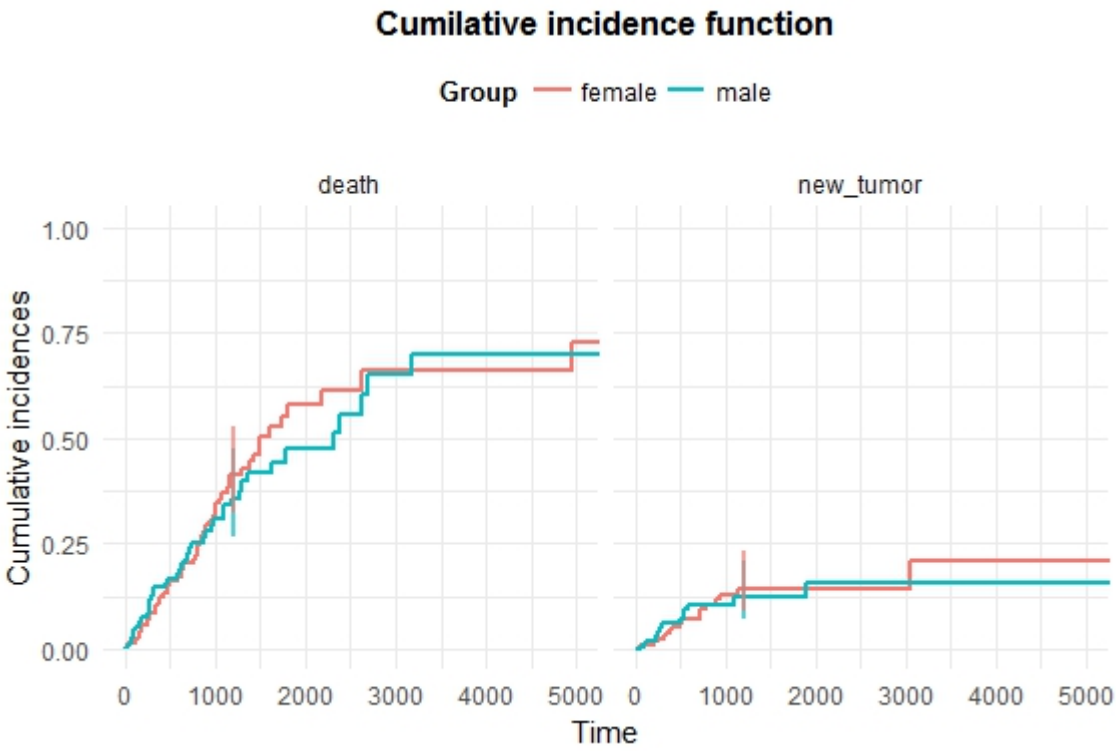
otrzymujemy wykres 2.4.

2.9. Testowanie modeli ryzyk konkurujących - funkcja testCuminc

Wyniki *testu dla K prób* (patrz: rozdział 1.7) otrzymujemy za pomocą funkcji `testCuminc`, której argumentem jest obiekt `fitCuminc`. Po wywołaniu:

```
testCuminc(ci = fitC)
```

otrzymujemy tabelkę przedstawioną w 2.5



Rysunek 2.4: Wynik funkcji `plotCuminc` na danych LUAD. Wykres przedstawia krzywe skumulowanych częstości dla pacjentów u których pierwszym zdarzeniem był zgon (*death*) oraz u których pierwszym zdarzeniem było wystąpienie nowego nowotworu (*new_tumor*). Krzywe są rysowane osobno dla kobiet (*female*) oraz dla mężczyzn (*male*).

	death	new_tumor
K-Sample Test	0.8278	0.9797

Rysunek 2.5: Wynik funkcji `testCuminc` na danych LUAD. W tabeli znajdują się p-wartości testu dla K-prób 1.77, badającego istotność występowania różnic w krzywych skumulowanych częstości dla kobiet i dla mężczyzn.

2.10. Estymacja modeli Coxa w przypadku występowania ryzyk konkurujących - funkcja fitReg

Ostatnią zaimplementowaną metodą jest model Coxa dla ryzyk konkurujących, nazywany także modelami regresji dla ryzyk konkurujących. Dopasowanie modelu dostajemy za pomocą funkcji `fitReg`, opartej na funkcji `crr` z pakietu `cmprsk`. Argumentem tej funkcji są ponownie **time**, **risk**, **group** oraz **cens**. Wartością tej funkcji jest lista, której elementami są dopasowania modelu Coxa dla ryzyk konkurencyjnych dla poszczególnych ryzyk i grup. W każdym przypadku dostajemy, w szczególności, następujące informacje:

- **coef**, wyestymowane parametry β (patrz: rozdział 1.8),
- **loglik**, logarytm pseudo-wiarygodności dla modelu,
- **score**, pochodna logarytmu pseudo-wiarygodności w punkcie **loglik**,
- **loglik.null**, logarytm pseudo-wiarygodności dla modelu zerowego (z wszystkimi parametrami równymi 0).

Dodatkowym, ostatnim elementem listy są wyniki *modyfikowanego testu LRT* (patrz: rozdział 1.8). Dla danych `LUAD`, po wywołaniu:

```
reg <- compRiskReg(time = LUAD$time,
                   risk = LUAD$event,
                   group = LUAD$gender,
                   data = LUAD,
                   cens = "alive")
```

dostajemy pięcioelementową listę, której pierwsze 4 elementy to dopasowania modeli Coxa w przypadku ryzyk konkurencyjnych dla poszczególnych grup i ryzyk. Piątym elementem jest tabelka z wynikami modyfikowanego testu ilorazy wiarygodności.

2.11. Testowanie modeli Coxa w przypadku występowania ryzyk konkurujących - funkcja testReg

Aby uzyskać tabelkę z p-wartościami dla modyfikowanego testu ilorazu wiarygodności, wykorzystujemy funkcje `testReg`, której argumentami jest objekt **fitReg** oraz **conf.int**. Wynikiem jest tabelka zawierająca p-wartości. Przykład dla danych `LUAD` został przedstawiony na rysunku 2.6.

```
testReg(fitReg = reg,
        conf.int = 0.95)
```

	death	new_tumor
CompRisk likelihood ratio test	0.82	0.91

Rysunek 2.6: Wynik funkcji `testReg` na danych LUAD. w tabeli znajdują się p-wartości modyfikowanego testu ilorazu wiarygodności dla modeli Coxa w przypadku występowania ryzyk konkurujących. Badana jest istotność występowania różnic w modelach dla kobiet i dla mężczyzn.

Number at risk

death

	0	1000	2000	3000	4000	5000
female	280	65	35	29	29	28
male	242	59	33	27	26	25

new_tumor

	0	1000	2000	3000	4000	5000
female	280	90	72	68	67	67
male	242	80	60	58	58	57

Rysunek 2.7: Wynik funkcji `riskTab`. Tabela przedstawiająca liczbę jednostek narażonych na ryzyko w czasie dla obydwu ryzyk w rozróżnieniu na płeć. Punkty, w których liczone są jednostki narażone na ryzyko odpowiadają punktom na osi poziomej wykresu z krzywymi przeżycia.

2.12. Zliczenia jednostek narażonych na ryzyko - funkcja riskTab

Tabelkę, w której znajduje się liczba jednostek narażonych na ryzyko w danych grupach można otrzymać za pomocą funkcji `riskTab`, której argumentami są `time`, `risk`, `group`, `cens`, zdefiniowane powyżej oraz argument `title` pozwalający na podanie tytułu tabelki (domyślnie: *Number at risk*). Dla danych `LUAD`, po wywołaniu:

```
riskTab(time =LUAD$time,
        risk = LUAD$event,
        group = LUAD$gender,
        cens = "alive",
        title = "Number␣at␣risk")
```

otrzymujemy tabelkę jak w 2.7.

Number of events

death

	0	1000	2000	3000	4000	5000
female	0	40	53	55	55	56
male	1	33	41	45	46	46

new_tumor

	0	1000	2000	3000	4000	5000
female	0	15	16	16	17	17
male	0	12	14	14	14	14

Rysunek 2.8: Wynik funkcji `eventTab`. Tabela przedstawiająca liczbę zdarzeń danego typu, które nastąpiły do danego czasu w rozróżnieniu na płeć. Punkty, w których liczone są wystąpienia zdarzeń odpowiadają punktom na osi poziomej wykresu z krzywymi skumulowanych częstości.

2.13. Zliczenia wystąpień zdarzeń - funkcja eventTab

Analogicznie, możemy otrzymać teraz tabelę zawierającą informację o liczbie zdarzeń do danego czasu t. Argumentami tej funkcji są ponownie `time`, `risk`, `group`, `data`, `cens`, zdefiniowane powyżej oraz argument `title` pozwalający na podanie tytułu tabelki (domyślnie: *Number of events*). Dla danych `LUAD`, po wywołaniu:

```
eventTab(time = LUAD$time,
         risk = LUAD$event,
         group = LUAD$gender,
         cens = "alive",
         title = "Number␣of␣events")
```

otrzymujemy tabelkę jak w 2.8.

2.14. Sumaryczny raport - funkcja summarizeCR

Aby możliwe ułatwić korzystanie z biblioteki `cr17`, zaimplementowana została funkcja `summarizeCR`, dzięki której można otrzymać wyniki z wszystkich opisanych powyżej funkcji w przejrzystym raporcie. Po wywołaniu:

```
summarizeCR(time = LUAD$time,
            risk = LUAD$event,
            group = LUAD$gender,
            cens = "alive",
            target = 1200,
            rho = 0,
            type = "kaplan-meier",
            conf.int = 0.95,
            conf.type = "log",
            ggtheme = theme_minimal(),
            titleSurv = "Survival␣curves",
            titleCuminc = "Cumulative␣incidence␣function",
            xtitle = "Time",
            ytitleSurv = "Probability␣of␣survivng␣up␣to␣time␣t",
            ytitleCuminc = "Cumulative␣incidences",
            legendtitle = "Group",
            riskTabTitle = "Number␣at␣risk",
            eventTabTitle = "Number␣of␣events")
```

otrzymujemy raport jak na rysunku 2.9. Jak łatwo zauważyć, raport można podzielić na dwie kolumny. Pierwsza z nich przedstawia rezultaty z modelowania analizy przeżycia dla ryzyk z osobna, podczas gdy druga z nich dotyczy ryzyk konkurujących. Punkty w czasie, dla których zostały wyliczone wartości w tabelce zawierającej liczbę jednostek narażonych na ryzyko odpowiadają tym na przedstawiających krzywe przeżycia. Analogicznie, punkty w czasie, dla których zliczone zostały wystąpienia zdarzeń odpowiadają osi poziomej wykresu w krzywymi skumulowanych częstości. Poprzez podanie odpowiednich argumentów można zmienić poziom ufności, dla których liczone są przedziały ufności i p-wartości, zmienić temat wykresów oraz dopasować własne tytuły wykresów i tabelek.

Rysunek 2.9: Wynik funkcji `summarizeCR` na danych LUAD. Sumaryczny raport przedstawiający porównanie czasów przeżycia w grupach dla poszczególnych ryzyk. W górnej części znajdują się dwa wykresy - wykres przedstawiający krzywe przeżycia (po lewej stronie) oraz wykres przedstawiający krzywe skumulowanych częstości (po prawej stronie). W środkowej części znajdują się tabelka ze zliczeniami jednostek narażonych na ryzyko (po lewej stronie) oraz tabelka ze zliczeniami zdarzeń (po prawej stronie). Na dole raportu znajdują się wyniki poszczególnych testów badających różnice w modelach pomiędzy grupami - w tym przypadku pomiędzy kobietami a mężczyznami. Z lewej strony znajduje się tabelka z p-wartościami dla testów opartych na analizie przeżycia, po prawej stronie natomiast przedstawione są wyniki testów dla modeli ryzyk konkurujących.

Rozdział 3

Przykład zastosowania na danych o pacjentach z nowotworami układu krwiotwórczego

3.1. Opis danych

Praktyczne zastosowanie pakietu przedstawię na danych pochodzących z projektu *InfAza*, będącego wspólnym przedsięwzięciem Warszawskiego Uniwersytetu Medycznego, pod przewodnictwem dra n. med. Krzysztofa Mądrego oraz lek. med. Karola Lisa wraz z Uniwersytetem Warszawskim pod przewodnictwem dra hab. Przemysława Biecka. Dane te, o roboczej nazwie `infaza`, zawierają informacje o pacjentach chorujących na jedną z trzech pokrewnych ze sobą chorób:

- **AML** (ang. *acute myeloid leukemia*) - ostrą białaczkę szpikową,
- **CMML** (ang. *chronic myelomonocytic lekuemia*) - przewlekłą białaczkę mielomonocytową,
- **MDS** (ang. *myelodysplastic syndrome*) - zespołem mielodysplastycznym, nazywanym także stanem przedbiałaczkowym.

Choroby te zaliczane są do nowotworów układu krwiotwórczego i powodują, między innymi, znacznie zmniejszoną odporność organizmu. Wszyscy pacjenci poddani zostali innowacyjnej terapii azacytadyną, w comiesięcznych cyklach podawania leku. Głównym celem gromadzenia danych było zdobycie wiedzy na temat ryzyka wystąpienia infekcji podczas 3 pierwszych miesięcy terapii oraz czynników wpływających na to ryzyko. Analiza statystyczna w ramach projektu, którą wykonywałam, obejmowała czyszczenie i eksplorację danych, budowanie modeli liniowych, testowanie istotności parametrów, stworzenie klasyfikacji pacjentów ze względu na ryzyko wystąpienia infekcji oraz wizualizacje.

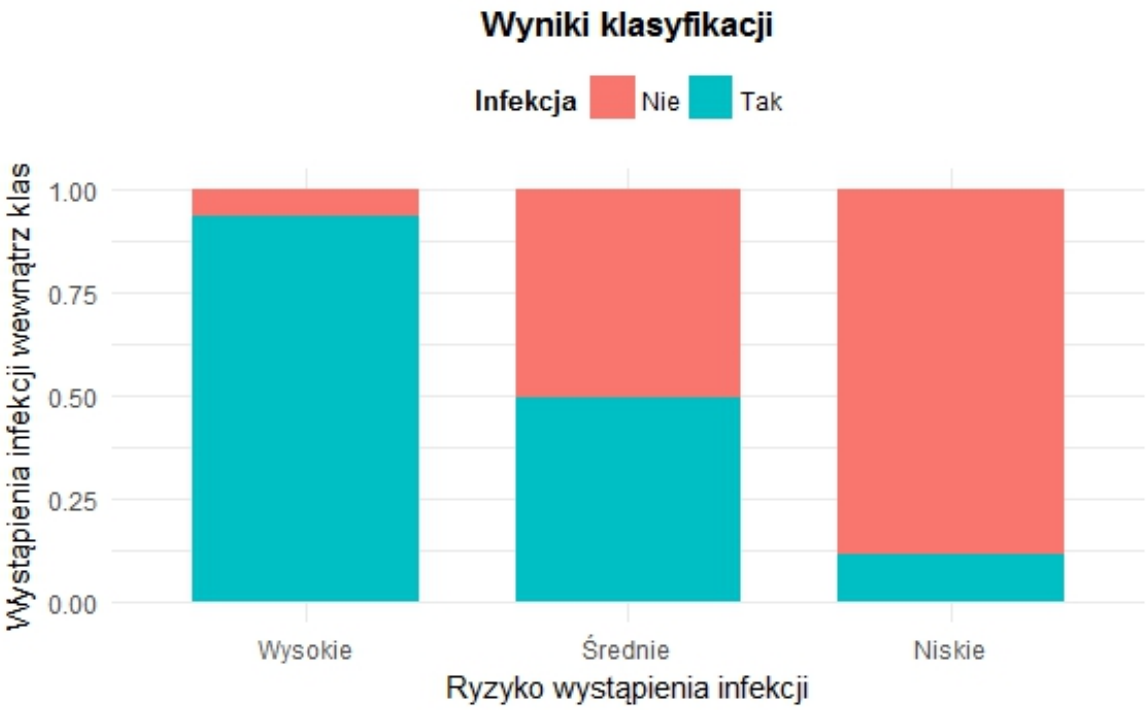
3.2. Eksploracja danych

Łącznie dane zawierają informacje o 298 osobach z 10 ośrodków medycznych w Polsce. Dla każdej obserwacji posiadamy następujące wiadomości:

- wiek i płeć pacjenta,
- ośrodek, w którym pacjent był leczony,

- rozpoznanie (AML, CMML, MDS),
- łączna liczba podanych cykli azacytadyny,
- czy u pacjenta nastąpiła infekcja, cykl przy którym wystąpiła infekcja, rodzaj infekcji (grzybicza, bakteryjna, wirusowa),
- czy u pacjenta zastosowana została profilaktyka przeciwwirusowa, przeciwbakteryjna lub przeciwgrzybicza,
- czy nastąpił zgon pacjenta,
- całkowity czas obserwacji (czas od rozpoczęcia leczenia azacytadyną do zgonu lub końca obserwacji),
- czas od diagnozy do rozpoczęcia leczenia,
- wyniki podstawowych badań wykonanych na początku leczenia (m.in. liczba limfocytów, neutrofili i monocytów, poziom ferrytyny, albuminy i kreatyniny, odstek blastów w szpiku),
- występowanie innych chorób (inny nowotwór, cukrzyca, niewydolność serca),
- różne klasyfikacje stanu zdrowia pacjenta (m.in. WHO).

W ramach projektu, najważniejszym zadaniem było wybranie modelu regresji logistycznej, w której zmienną objaśnianą była zmienna binarna oznaczająca wystąpienie infekcji w ciągu pierwszych 3-ech miesięcy leczenia Azacytadyną. Ważną cechą szukanego modelu była łatwość jego interpretacji, skąd ograniczaliśmy się do co najwyżej sześciu zmiennych objaśniających. Pozostałymi czynnikami wyboru modelu były wyniki testów na istotność parametrów, powierzchnia pod krzywą ROC (ang. *Receiver Operating Characteristic*) oraz dokładność (ang. *accuracy*). Z powodu dużej ilości brakujących wartości wykonana została imputacja danych (wyniki przedstawione w tej pracy zostały otrzymane na pierwotnych danych). W końcowym modelu znajdowały się następujące zmienne objaśniające: rozpoznanie, klasyfikacja WHO, zależność od przetoczeń krwi, poziom albumy oraz ilość neutrofili we krwi oraz procentowa zawartość blastów w szpiku. Na podstawie wybranego modelu, stworzona została 3-stopniowa klasyfikacja, dzięki której możliwy jest podział pacjentów ze względu na poziom zagrożenia wystąpienia infekcji. Wyniki z otrzymanej klasyfikacji zostały przedstawione na rysunku 3.1.

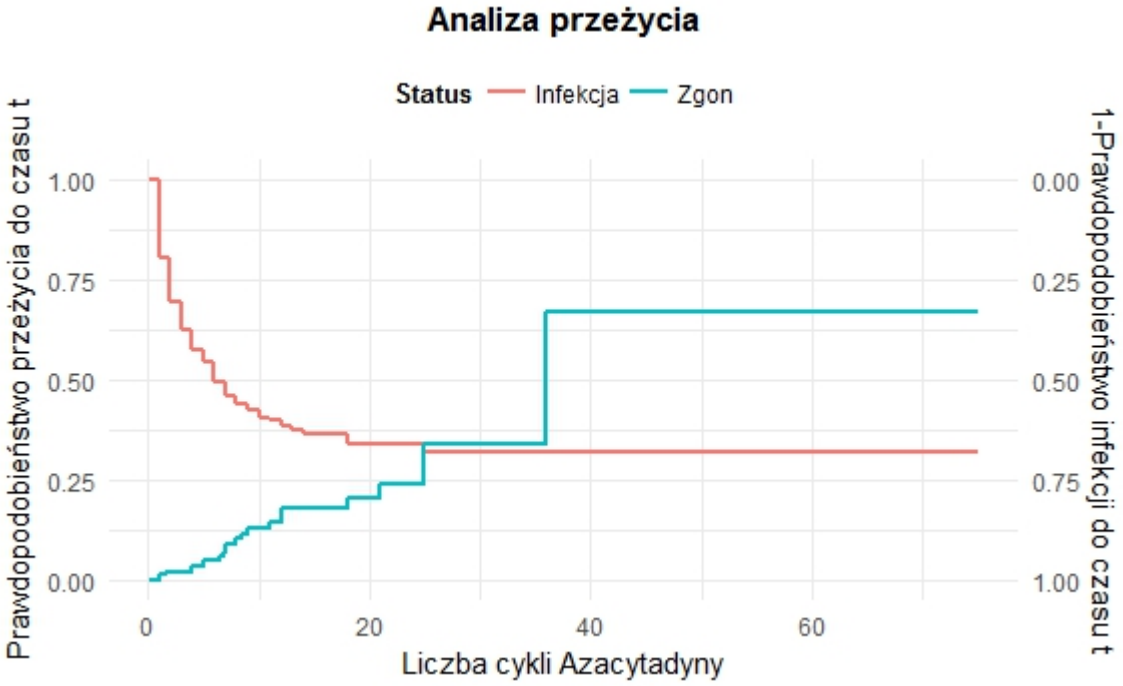


Rysunek 3.1: Wyniki 3-stopniowej klasyfikacji zastosowanej na danych `infaza`, powstałej w ramach projektu InfAza. Wykres przedstawia procentowy udział pacjentów u których wystąpiła infekcja wśród wszystkich pacjentów zakwalifikowanych do danej klasy. Jak wynika z wykresu, spośród wszystkich pacjentów, u których stwierdzono wysokie zagrożenie wystąpienia infekcji, u około 95% nastąpiła nastąpiła. Spośród pacjentów o średnim ryzyku, zdarzenie wystąpiło w około 50% przypadków, natomiast wśród pacjentów o niskim zagrożeniu, infekcja wystąpiła tylko w około 10% przypadków.

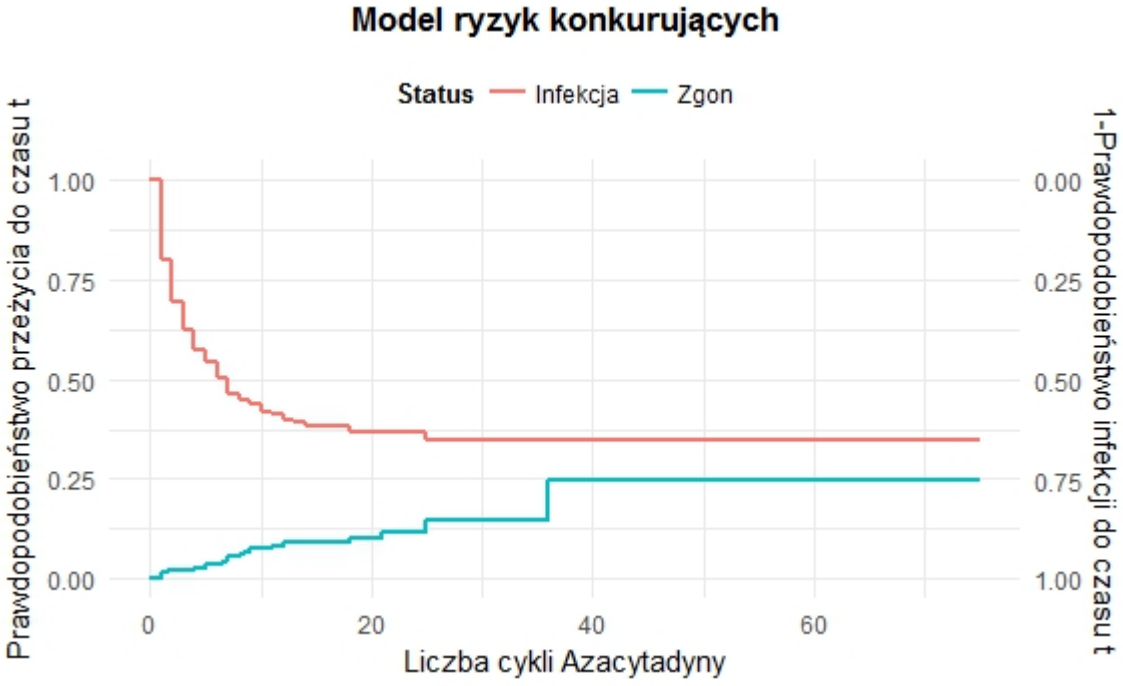
3.3. Analiza przeżycia a modele ryzyk konkurujących

W poprzednich rozdziałach zostało podkreślone, że stosowanie analizy przeżycia w przypadku występowania ryzyk konkurencyjnych może prowadzić do błędnych wyników. Przykład takiej sytuacji zaprezentuję na danych `infaza`. Na rysunku 3.2 przedstawione zostały krzywe przeżycia Kaplana-Meiera dla obydwu ryzyk, estymowane niezależnie (to znaczy, w każdym przypadku rozpatrywane były tylko zdarzenia danego rodzaju). Obserwacje, u których wystąpiło zdarzenie drugiego rodzaju traktowane były jako cenzorowane. W celu lepszego zwizualizowania problemu, krzywe dla obydwu ryzyk zostały przedstawione na osiach o przeciwnym kierunku wzrastania. Przecinające się osie świadczą o tym, że dla odpowiednio dużych czasów, prawdopodobieństwo wystąpienia jednego z dwóch ryzyk jest większe od 1.

Natomiast w przypadku zastosowania modelu ryzyk konkurujących, oba wykresy nie przecinają się, co widać na rysunk 3.3. Nie otrzymujemy tutaj prawdopodobieństwa wykraczającego poza przedział $[0, 1]$.



Rysunek 3.2: Krzywe przeżycia w przypadku zastosowania analizy przeżycia do obydwu ryzyk niezależnie. Przecinające sie krzywe świadczą o tym, że prawdopodobieństwo wystąpienia jednego ze zdarzeń od pewnego momentu jest większe niż 1.



Rysunek 3.3: Krzywe skumulowanych częstości w przypadku zastosowania modeli ryzyk konkurujących. Krzywe nie przecinają się, nie otrzymujemy prawdopodobieństwa wystąpienia zdarzenia poza przedziałem $[0, 1]$.

3.4. Zastosowanie biblioteki cr17

W analizie skupimy się na dwóch zmiennych grupujących, które, w wyniku pracy nad projektem *InfAza* okazały się mieć istotny wpływ na wystąpienie infekcji. Pierwszą z nich będzie `Rozpoznanie`. Będziemy chcieli sprawdzić, czy ryzyko wystąpienia infekcji i zgonu różni się w zależności od typu choroby. Drugim wyborem będzie zmienna `WHO`, określająca stan zdrowia pacjenta według *Międzynarodowej Klasyfikacji Funkcjonowania, Niepełnosprawności i Zdrowia ICF* (ang. *International Classification of Functioning, Disability and Health*) [4]. Klasyfikacja ta opiera się na szerokim spektrum informacji o danym pacjencie - zarówno o jego stanie zdrowia fizycznym i psychicznym, jak i o czynnikach indywidualnych (np. wiek, płeć) i środowiskowych (np. miejsce pracy). Na potrzeby projektu *Infaza* wynik klasyfikacji został zakodowany jako zmienna binarna, dla której 0 oznacza *niski* poziom stanu zdrowia pacjenta, zaś 1 *wysoki*. W tabeli 3.1 przedstawione zostały zliczenia obserwacji według podziału na rozpoznanie, a w tabeli 3.2 zliczenia obserwacji według podziału na stan klasyfikację WHO.

Tablica 3.1: Tabela liczebności według rozpoznania dla danych Infaza.

	Infekcja	*Zgon*	*Brak zdarzenia*	Razem
AML	48	7	16	71
CMML	15	4	14	33
MDS	94	21	63	178
Razem	157	32	93	282

Tablica 3.2: Tabela liczebności według klasyfikacji ICF dla danych Infaza.

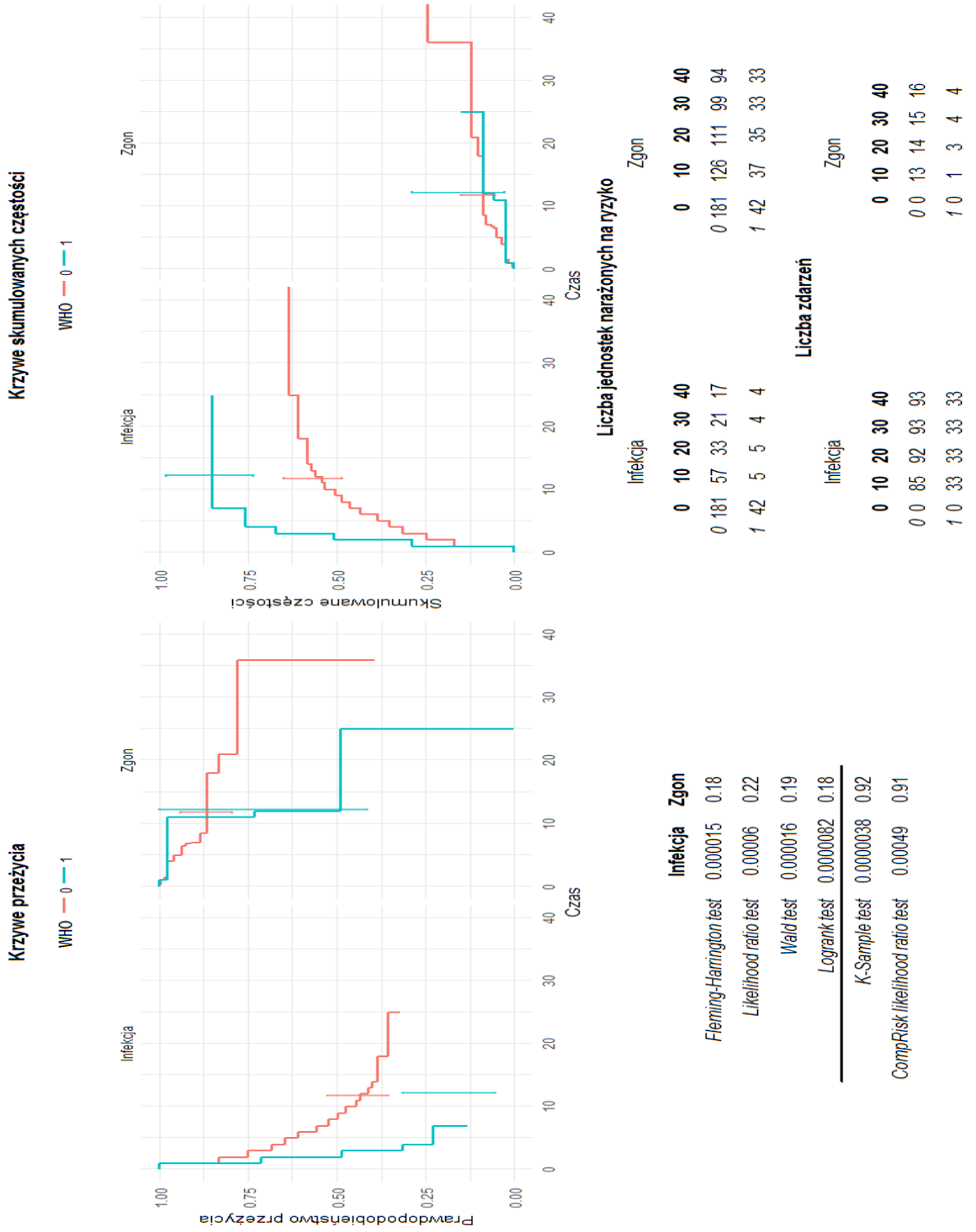
	Infekcja	*Zgon*	*Brak zdarzenia*	Razem
0	101	21	76	198
1	34	5	5	44
Razem	135	26	81	242

Przyjrzyjmy się teraz raportom wygenerowanym za pomocą funkcji `summarizeCR` z pakietu `cr17`. Na rysunku 3.4 przedstawione zostały wyniki dla podziału na grupy ze względu na rozpoznanie. Jak widzimy, na poziomie istotności 95%, testy oparte na metodach analizy przeżycia sugerują, że ryzyko wystąpienia zgonu jest istotnie różne pomiędzy grupami (p-wartość 0.05). Podczas gdy metody oparte na modelowaniu ryzyk konkurencyjnych nie wskazują na istnienie takich różnic (p-wartość 0.05). W przypadku zdarzenia, jakim jest infekcja, oba podejścia sugerują brak różnic między grupami.

W przypadku podziału ze względu na stan zdrowia pacjenta według klasyfikacji WHO 3.5, oba podejścia sugerują brak różnic w ryzyku wystąpienia zgonu. Metody oparte na metodach analizy przeżycia odrzucają hipotezę zerową, mówiącą o braku różnic między grupami w przypadku zgonu. Testy oparte na modelowaniu ryzyk konkurencyjnych dają natomiast różne wyniki. Test dla K prób *CompRiskLRT*, sugeruje brak różnic (p-wartość > 0.05), modyfikowany test LRT sugeruje natomiast występowanie istotnej różnicy.

Wcześnie zostało zauważone, że stosowanie metod analizy przeżycia dla poszczególnych zdarzeń nie jest adekwatne do modelowania w przypadku występowania ryzyk konkurencyjnych. Przedstawione tutaj wyniki wskazują, iż podejście to nie jest adekwatne także do testowania różnic między grupami.

Rysunek 3.4: Wyniki funkcji summarizeCR zastosowany do danych `infaza` dla zmiennej grupującej `Rozpoznanie`. Testy oparte na analizie przeżycia sugerują istotność różnic między danymi podtypami choroby w przypadku zgonu i infekcji na poziomie istotności 95%. Testy oparte na modelowaniu ryzyk konkurujących sugerują występowanie tych różnic jedynie w przypadku infekcji.



Rysunek 3.5: Wynki funkcji summarizeCR zastosowany do danych `infaza` dla zmiennej grupującej `WHO`. Ponownie, testy oparte na analizie przeżycia sugerują istotność różnic wśród pacjentów o różnej klasyfikacji WHO w przypadku zgonu i infekcji na poziomie istotności 95%. Testy oparte na modelowaniu ryzyk konkurujących sugerują występowanie tych różnic jedynie w przypadku infekcji.

Zakończenie

Powyższa praca stanowi opis biblioteki `cr17`, którą przygotowałam, pod przewodnictwem dra hab. Przemysława Biecka, w ramach projektu *Infaza*. Opisane zostało także jej podłoże merytoryczne. Pakiet umożliwia zarówno diagnostykę, jak i wizualizację modeli analizy przeżycia i ryzyk konkurujących, z szczególnym naciskiem na testowanie różnic pomiędzy grupami. W pracy zaprezentowane zostały także przykłady, które uwidaczniają problemy powstające przy użyciu modeli analizy przeżycia w przypadku istnienia ryzyk konkurencyjnych. Funkcjonalność pakietu zaprezentowana została na przykładzie danych medycznych, wykorzystywanych podczas projektu *Infaza*.

Na chwilę obecną udostępniona została pierwsza wersja pakietu. W dalszych krokach planowane są dalsze testy oraz rozszerzenia. Jednym z pomysłów jest umożliwienie modelowania parametrycznego dla ryzyk konkurujących. Dla takich modeli można testować równość parametrów w rozkładach dla poszczególnych grup. Mając estymacje modeli w poszczególnych grupach można testować różnice między nim za pomocą ogólnych metod analizy statystycznej.

Wierzymy, że biblioteka `cr17` będzie wykorzystywana w szerokim spektrum zastosowań i z biegiem czasu uda nam się stworzyć kompleksowe narzędzie do analizy modeli ryzyk konkurujących.

Bibliografia

[1] O. Aalen, *Nonparametric estimation of partial transition probabilities in multiple decrement models*, Ann. Statist. 6 534-545, 1978.

[2] O. Aalen, *Nonparametric inference for a family of counting processes*, 1978.

[3] A. Allignol, J. Beyersmann, M.Schumacher, *Competing Risks and Multistate Models with R*, Springer, 2012.

[4] A. Wilmowska-Pietruszyńska, D. Bilski, *Międzynarodowa Klasyfikacja Funkcjonowania, Niepełnosprawności i Zdrowia*, https://www.pfron.org.pl/kn/poprzednie-numery/181,Miedzynarodowa-Klasyfikacja-Funkcjonowania-Niepelnosprawnosci-i-Zdrowia-Internat.html, dostępny 30.06.2017.

[5] D.Cox, D.Oakes, *Analysis of Survival Data*, Chapman and Hall/CRC, London, New York, 1984.

[6] J. Fine, R. Grey *A Proportional Hazards Models for the Subdistribution of a Competing Risk*, Journal of the American Statistical Association, Vol. 94, No. 446, 1999.

[7] T. Fleming, D. Harrington, *Nonparametric estimation of the survival distribution in censored data*,Comm. in Statistics 13, 2469-86, 1983.

[8] B. Gray, *cmprsk* Package, https://cran.r-project.org/web/packages/cmprsk/cmprsk.pdf, dostępny 16.07.2017.

[9] R. Grey, *A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk*, The Annals of Statistics, Vol. 16, No.3, pp. 1141-1154, 1988.

[10] S. Johansen, *The product limit estimator as a maximum likelihood estimator*, Scad. J. Statist. 5 195-199, 1978.

[11] M. Kosiński, *RTCGA.clinical: Clinical datasets from The Cancer Genome Atlas Project*, R package version 20151101.6.0.m, 2016.

[12] D. Kleinbaum, M. Klein, *Survival Analysis*, Springer, 2012.

[13] D. Moore, *Applied Survival Analysis Using R*, Springer, 2016.

[14] M. Pešta *http://www.karlin.mff.cuni.cz/pesta/NMFM404/survival.html*, Charles University, dostępny 28.06.2017.

[15] S. Sawyer, *The Greenwood and Exponential Greenwood Confidence Intervals in Survival Analysis*, 2003.

[16] T. Therneau, *Survival* Package Manual, https://cran.r-project.org/web/packages/survival/index.html, dostępny 16.07.2017.