Politechnika Wrocławska

Wydział Matematyki
Kierunek studiów: Matematyka
Specjalność: Statystyka i analiza danych

Praca dyplomowa – magisterska

MODELE NIEPROPORCJONALNYCH HAZARDÓW

Monika Mrozek

słowa kluczowe:
analiza przeżycia, model proporcjonalnych hazardów Coxa, modele nieproporcjonalnych hazardów, interakcje pomiędzy zmienną a czasem, analiza danych

krótkie streszczenie:
W pracy przedstawiono model proporcjonalnych hazardów Coxa oraz modele nieproporcjonalnych hazardów wraz z przykładem praktycznego wykorzystania wybranych modeli. W części teoretycznej zaprezentowano w szczególności różne metody estymacji współczynników opisywanych modeli. Część praktyczna dotycząca analizy danych rzeczywistych została przeprowadzona w pakiecie R. Poprzedzono ją opisem teoretycznym wybranych narzędzi i metod.

Opiekun pracy dyplomowej	Dr hab. Alicja Jokiel-Rokita	.....	.....
	Tytuł/stopień naukowy/imię i nazwisko	ocena	podpis

*Do celów archiwalnych pracę dyplomową zakwalifikowano do:**

a) kategorii A (akta wieczyste)

b) kategorii BE 50 (po 50 latach podlegające ekspertyzie)

** niepotrzebne skreślić*

pieczątka wydziałowa

Wrocław, rok 2021

Wrocław University
of Science and Technology

Faculty of Pure and Applied Mathematics
Field of study: Mathematics
Specialty: Statistics and Data Analysis

Master's Thesis

NON-PROPORTIONAL HAZARDS MODELS

Monika Mrozek

keywords:
survival analysis, Cox's proportional hazards model, non-proportional hazards models, covariate-time interactions, data analysis

short summary:
The thesis presents Cox's proportional hazards model and non-proportional hazards models together with the example of practical application of selected models. The theoretical part presents in particular various methods estimating of parameters of models described. The practical part was conducted in R package. It is preceded by the theoretical description of selected tools and methods.

Supervisor	Dr hab. Alicja Jokiel-Rokita	.....	.....
	Title/degree/name and surname	grade	signature

*For the purposes of archival thesis qualified to:**

a) category A (perpetual files)

b) category BE 50 (subject to expertise after 50 years)

** delete as appropriate*

stamp of the faculty

Wrocław, 2021

Spis treści

Spis rysunków

Spis tablic

Wstęp

Tematem niniejszej pracy są modele nieproporcjonalnych hazardów, które są głównie wykorzystywane w analizie przeżycia. Analiza przeżycia jest dziedziną statystyki badającą różne zjawiska i procesy ze szczególnym uwzględnieniem czasu przeżycia jako zmiennej głównego zainteresowania. Czas przeżycia to czas do wystąpienia danego zdarzenia bądź zdarzeń. Takim zdarzeniem może być śmierć pacjenta, nawrót choroby nowotoworowej czy awaria urządzenia. Jednym z głównym celów pracy jest opis modeli, które mogą być zastosowane, gdy hazard jest nieproporcjonalny, tzn., gdy ryzyko względne będące ilorazem odpowiednich funkcji hazardu zmienia się w czasie. Przedstawiliśmy także model proporcjonalnych hazardów Coxa, który do chwili obecnej jest najbardziej popularnym narzędziem, lecz znajdującym zastosowanie tylko w przypadku, gdy hazard jest proporcjonalny. W pracy przeprowadziliśmy także analizę danych rzeczywistych przy wykorzystaniu wybranych modeli.

W Rozdziale 1 zaprezentowaliśmy podstawowe pojęcia analizy przeżycia. W szczególności zdefiniowaliśmy pojęcie hazardu oraz ryzyka względnego, którego wartość w przypadku, gdy założenie proporcjonalności hazardów nie jest spełnione, zmienia się w czasie. W Rozdziale 1 przedstawiliśmy również niezbędne oznaczenia, które są wykorzystywane w dalszej części pracy.

Rozdział 2 zawiera opis modelu proporcjonalnych hazardów Coxa oraz modeli nieproporcjonalnych hazardów, które są jego modyfikacjami. Pierwszym modelem nieproporcjonalnych hazardów jest model ze współczynnikiem zależnym od czasu. Przedstawiliśmy również model Yanga-Prentice'a-Diao-Zenga. Z powodu trudności, które pojawiały się w jego interpretacji zaproponowaliśmy model XS, który wykorzystuje interakcje pomiędzy zmienną a czasem. Dla każdego z modeli opisaliśmy ideę estymacji jego współczynników.

W Rozdziale 3 przedstawiliśmy pozostałe modele nieproporcjonalnych hazardów, które stanowią alternatywę dla modelu proporcjonalnych hazardów Coxa. Jednym z najczęściej wykorzystywanych modeli w takiej sytuacji jest model proporcjonalnych szans. W kolejnym podrozdziale opisaliśmy klasę modeli transformacji, których szczególnymi przypadkami są model proporcjonalnych hazardów Coxa i model proporcjonalnych szans. Przedstawiliśmy także model ryzyka addytywnego ze szczególnym uwzględnieniem podobieństw i różnic między tymże modelem a modelem proporcjonalnych hazardów Coxa. Na koniec opisaliśmy model przyspieszonego czasu awarii, który posiada wiele zalet w porównaniu do modelu ryzyka addytywnego i znajduje szerokie zastosowanie m. in. w analizie niezawodności. Dla modeli przedstawionych w Rozdziale 3 również przedstawiliśmy ideę estymacji wektora współczynników.

W Rozdziale 4 zaproponowaliśmy ogólny sposób modyfikacji podstawowych wersji przedstawionych modeli dla danych z dużą liczbą zer, tzn. danych, w których pojawia się zerowy czas przeżycia.

Rozdział 5 stanowi wstęp do części praktycznej pracy. Opisaliśmy w nim metody i narzędzia diagostyczne, które są pomocne we właściwym w pewnym sensie wyborze

modelu oraz weryfikacji jakości jego dopasowania.

Część praktyczną pracy przedstawiliśmy w Rozdziale 6. Analiza danych rzeczywistych została przeprowadzona dla wybranych modeli za pomocą języka programowania *R*. Dane, które zostały przez nas wykorzystane można znaleźć w bibliotekach `dynpred`, `timereg` oraz `mfp`. Dopasowane modele to model proporcjonalnych hazardów Coxa, model ze współczynnikiem zależnym od czasu, model proporcjonalnych szans oraz model ryzyka adddytywnego. Budowę tychże modeli umożliwiły kolejno funkcje `coxph()` z pakietu `survival`, `timecox()` z pakietu `timereg`, `prop.odds()` z wcześniej wymienionego pakietu oraz `ahaz()` z pakietu `ahaz`.

Rozdział 1

Wprowadzenie

Na początku niniejszej pracy zostanie przedstawione pojęcie analizy przeżycia oraz główne obiekty zainteresowania i badań tej dziedziny nauki. Dodatkowo zaprezentujemy najważniejsze pojęcia, które będą pojawiać się w przeprowadzanych rozważaniach.

Analiza przeżycia jest to dziedzina statystyki zajmująca się badaniem różnych procesów za pomocą metod oraz narzędzi statystycznych. Zmienną głównego zainteresowania jest czas przeżycia. Jest to czas, który upłynie do wystąpienia wcześniej określonego zdarzenia bądź zdarzeń. Takimi zdarzeniami mogą być śmierć pacjenta, awaria systemu lub urządzenia czy zużycie produktu. Problematyka badań analizy przeżycia pojawia się także w innych gałęziach nauki takich jak ekonometria czy socjologia. Widać zatem, że obszar możliwych zastosowań jest bardzo rozległy. Co więcej, analiza przeżycia jest nauką, której popularność nie słabnie od wielu lat o czym świadczy mnogość wydanych publikacji, przeprowadzanych analiz, a także gotowych funkcji zaimplementowanych w różnych pakietach statystycznych.

Dzięki analizie przeżycia można odpowiedzieć na wiele ważnych i interesujących pytań. Przykładowo, przy użyciu metod i narzędzi, które oferuje nam ta gałąź statystyki, istnieje możliwość oszacowania frakcji populacji, która przeżyje określony okres czasu. Metody analizy przeżycia pomagają także określić w jaki sposób cechy człowieka oraz warunki życia zwiększają bądź zmniejszają prawdopodobieństwo jego śmierci w określonym czasie. Informacje te są niewątpliwie bardzo cenne nie tylko dla branży medycznej, lecz także dla wielu firm, bowiem otrzymane prognozy znacząco wpływają na strategie ich postępowania.

W analizie przeżycia pojawia się wiele pojęć, których znajomość jest niezbędna do prawidłowego wykonania oraz interpretacji wyników różnego rodzaju analiz. W niniejszym rozdziale przedstawimy te z nich, które pojawiają się przy opisie modeli.

Termin, który jest nieodłącznie związany z analizą przeżycia to funkcja przeżycia (ang. *survival function*). Niech T będzie nieujemną zmienną losową odpowiadającą za czas do wystąpienia zdarzenia, pochodzącą z rozkładu o dystrybuancie F. Co więcej, zakładamy, że F jest różniczkowalna dla $t > 0$ i spełnia warunek $F(0) = 0$. Z tego powodu oraz zależności pomiędzy dystrybuantą a gęstością dla $t > 0$ zachodzi $f(t) = F'(t)$. Poniżej przedstawiamy definicję funkcji przeżycia.

Definicja 1.1. Funkcją przeżycia S (ang. *survival function*) nazywamy funkcję postaci

$$S(t) = P(T > t).$$

Pomiędzy funkcją przeżycia S, dystrybuantą F oraz gęstością f zachodzą związki

$$F(t) = P(T \leqslant t) = 1 - S(t)$$

oraz

$$f(t) = F'(t) = -S'(t).$$

Warto zaznaczyć, że funkcja przeżycia nie może być dowolną funkcją, ponieważ powinna spełniać pewne ściśle określone warunki. Przy powyższych założeniach wartość tej funkcji w zerze wynosi jeden, tzn. $S(0) = 1$, ponieważ w chwili, gdy rozpoczynamy analizę, zdarzenie polegające na tym, że obserwowana jednostka przeżyje jest zdarzeniem pewnym. Ponadto wymaga się, aby funkcja przeżycia była nierosnąca, co formalnie oznacza, że $S(u) \leqslant S(t)$, gdy $u \geqslant t$. Jest to założenie znajdujące uzasadnienie w szczególności wobec faktu przyjmowania wartości jeden w zerze – prawdopodobieństwo, które jest wartością funkcji przeżycia, nie może przekroczyć jej wartości początkowej. Ponadto zauważmy, że przy założeniu $u \geqslant t$ z warunku $T > u$ wynika natychmiast nierówność $T > t$. Innymi słowy jeśli wiadomym jest, że czas przeżycia był większy od u, to przy założeniu $u \geqslant t$ czas przeżycia będzie z pewnością większy od t. To założenie gwarantuje osiągnięcie przez jednostkę starszego wieku pod warunkiem, że przeżyła ona lata wcześniejsze. Przyjmujemy również, że $\lim_{n\to\infty} S(t) = 0$, ponieważ w większości przypadków czas życia nie jest nieograniczony. Czasem jednak zdarza się, że założenie to jest pomijane. Dotyczy to przede wszystkim zastosowań w naukach chemicznych, gdzie pewnego rodzaju izotopy mogą trwać w nieskończoność. Wtedy warunek osiągania granicy równej zero w nieskończoności nie jest nakładany.

Kolejnym ważnym pojęciem, które dodatkowo pojawia się w tytule niniejszej pracy jest hazard. Poniżej przedstawiamy definicje funkcji hazardu (ang. *hazard function*) i funkcji skumulowanego hazardu (ang. *cumulative hazard function*).

Definicja 1.2. Funkcją hazardu (ang. *hazard function*) nazywamy funkcję postaci

$$h(t) = \lim_{\Delta t\to 0^+} \frac{P(T \leqslant t + \Delta t | T > t)}{\Delta t}.$$

Dla funkcji hazardu zachodzi następujący fakt.

Fakt 1.3. *Przy przyjętych założeniach*

$$h(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} \tag{1.1}$$

Wartą uzasadnienia jest pierwsza równość wyrażenia (1.1). Poniżej zaprezentowaliśmy jej dowód.

Dowód.

$$\begin{aligned} h(t) &= \lim_{\Delta t\to 0^+} \frac{P(T \leqslant t + \Delta t | T > t)}{\Delta t} = \lim_{\Delta t\to 0^+} \frac{P(t < T \leqslant t + \Delta t)}{\Delta t P(T > t)} = \\ &= \lim_{\Delta t\to 0^+} \frac{F(t + \Delta t) - F(t)}{\Delta t S(t)} = \frac{f(t)}{S(t)}. \end{aligned}$$

□

Po zdefiniowaniu funkcji hazardu, możemy przejść do pojęcia funkcji skumulowanego hazardu.

Definicja 1.4. Funkcją skumulowanego hazardu (ang. *cumulative hazard function*) nazywamy funkcję postaci

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

gdzie h jest funkcją hazardu.

Ponadto z Faktu 1.3 otrzymujemy $H(t) = -\ln S(t)$.

W rozważaniach często będziemy także używać pojęcia ryzyka względnego (ang. *hazard ratio*). Niech x_i będzie wektorem charaktetrystyk, czyli wartości zmiennych objaśniających, związanych z i-tą jednostką. Wtedy przez x_1 oznaczamy wektor charakterystyk dla pierwszej jednostki, natomiast przez x_2 oznaczamy wektor charakterystyk dla drugiej jednostki. Ponadto przez $h(t|x_1)$ oraz $h(t|x_2)$ oznaczamy hazard warunkowy odpowiednio dla jednostki pierwszej oraz drugiej.

Definicja 1.5. Ryzykiem względnym (ang. *hazard ratio*) nazywamy stosunek hazardów warunkowych, tzn.

$$\frac{h(t|x_1)}{h(t|x_2)}.$$

Z pojęciem ryzyka względnego związane jest zagadnienie proporcjonalności hazardów. Proporcjonalność ta ma miejsce, gdy ryzyko względne jest stałe. W szczególności, gdy założenie proporcjonalności hazardów jest spełnione, ryzyko względne nie może zależeć od czasu.

Po zdefiniowaniu ryzyka względnego możemy przejść do prezentacji pojęć krótkoterminowego i długoterminowego ryzyka względnego.

Definicja 1.6. Krótkoterminowym ryzykiem względnym (ang. *short-term hazard ratio*) nazywamy granicę postaci

$$\lim_{t\to 0} \frac{h(t|x_1)}{h(t|x_2)}.$$

Analogicznie definiujemy długoterminowe ryzyko względne.

Definicja 1.7. Długoterminowym ryzykiem względnym (ang. *long-term hazard ratio*) nazywamy granicę postaci

$$\lim_{t\to \infty} \frac{h(t|x_1)}{h(t|x_2)}.$$

Kolejnym ważnym terminem, który stale pojawia się przy prognozie czasu przeżycia za pomocą różnych modeli nieproporcjonalnych hazardów są dane cenzurowane. Są to dane, dla których pełna informacja o czasie przeżycia nie jest dostępna. Powody takich braków mogą być różne – jednym z nich może być zaniechanie przeprowadzania dalszych badań. Co więcej, istnieją różne typy cenzurowania. Jeden ze sposobów klasyfikacji danych cenzurowanych wyróżnia cenzurowanie prawostronne, lewostronne oraz przedziałowe. W cenzurowaniu prawostronnym zdarzenie występuje po zakończeniu prowadzenia obserwacji. W konsekwencji obserwowany jest tylko moment czasowy, w którym utraciliśmy dalszą informację na temat czasu wystąpienia zdarzenia. Cenzurowanie lewostronne definiuje się analogicznie, co oznacza, że zdarzenie występuje przed objęciem jednostki badaniem. W cenzurowaniu przedziałowym natomiast, ponownie nie mamy informacji na temat czasu wystąpienia zdarzenia, ponieważ może mieć ono miejsce przed czasem rozpoczęcia badania

lub po jego zakończeniu.

W niniejszej pracy będą rozpatrywane dane cenzurowane prawostronnie. Wiąże się to z koniecznością wprowadzenia pewnych oznaczeń. Niech T będzie zmienną losową związaną z czasem do wystąpienia zdarzenia, natomiast C zmienną losową związaną z czasem cenzurowania. Co więcej zakładamy, że zmienne te są niezależne. Dla każdej jednostki obserwujemy dane $T_i^* = \min(T_i, C_i)$ oraz $\delta_i = I(T_i \leqslant C_i)$, gdzie $I(\cdot)$ jest indykatorem. Ponadto przez $Y_i(t)$ będziemy oznaczać wartość związaną ze zmienną binarną, która pozwala stwierdzić czy i-ta obserwowana jednostka jest zagrożona, tzn. $Y_i(t) = I(t \leqslant T_i^*)$, natomiast przez $N_i(t)$ proces liczący postaci $N_i(t) = I(T_i^* \leqslant t, \delta_i = 1)$.

Rozdział 2

Model proporcjonalnych hazardów Coxa i jego modyfikacje

W niniejszym rozdziale prezentujemy model proporcjonalnych hazardów Coxa wraz z wybranymi jego modyfikacjami. Modyfikacje te są doskonałym narzędziem statystycznym w przypadku, gdy ryzyko względne nie spełnia warunku proporcjonalności. Model proporcjonalnych hazardów Coxa [1] stoi w opozycji do pozostałych modeli. Był on często zakładany w analizach, lecz niestety nie ma zastosowania w wielu sytuacjach praktycznych. Skłoniło to do podjęcia rozważań na temat odpowiednich jego modyfikacji. Opisywane modele prezentujemy dla przypadku, gdy dane są cenzurowane prawostronnie. Przedstawiliśmy również zarys problemu estymacji współczynników.

Pierwszy z opisanych modeli, który można wykorzystać w przypadku, gdy hazard jest nieproporcjonalny, jest modyfikacją modelu proporcjonalnych hazardów Coxa [2], która dopuszcza zmienność współczynników modelu w czasie. Kolejny z opisanych modeli został zaproponowany przez Yanga, Prentice'a, Diao i Zenga [3]. Ma on jednak niestety pewne wady. Pojawia się w nim bowiem trudność związana z określeniem relacji pomiędzy ryzykiem względnym a wektorem charakterystyk. Co więcej, pojawiają się trudności w interpretacji otrzymanych wyników. Ostatni z przedstawionych modeli, zaproponowany w [4], jest modyfikacją modelu Coxa, która obejmuje wprowadzenie interakcji pomiędzy zmienną a czasem. Nie ma on już wad modelu poprzedniego. Ponadto w opinii autorów artykułu, w którym został on zaproponowany, jest on modelem o najlepszych własnościach. Ryzyko względne przy założeniach tego modelu jest bowiem prostsze w interpretacji.

2.1 Model proporcjonalnych hazardów Coxa

Model proporcjonalnych hazardów dużym zainteresowaniem cieszył się już w roku 1972 po publikacji jednej z prac brytyjskiego statystyka Davida Coxa [1]. Naukowiec ten słynie głównie z powodu wspomnianego wcześniej modelu, którego jest twórcą. David Cox działa także prężnie w różnych innych dziedzinach statystyki. Obiektem jego zainteresowania są zagadnienia związane z m. in. regresją logistyczną [5]. Od jego nazwiska pochodzi również nazwa jednego z procesów punktowych będącego uogólnieniem procesu Poissona. Od momentu publikacji pracy Coxa dotyczącej modelu proporcjonalnych hazardów, model ten stał się popularnym narzędziem używanym w analizie danych cenzurowanych [4].

Niech x będzie wektorem charakterystyk, in. zmiennych objaśniających, który jest związany z i-tą obserwacją. Wtedy model proporcjonalnych hazardów Coxa przyjmuje

postać

$$h(t|x) = e^{\beta^T x} h_0(t), \tag{2.1}$$

gdzie β jest wektorem współczynników modelu, a $h_0(t) = h(t|x = 0)$ jest funkcją hazardu bazowego. Wyrażenie $e^{\beta^T x}$ nazywane jest czynnikiem multiplikatywnym. Nazwa ta związana jest z multiplikatywnym wpływem predyktorów na hazard. W przypadku funkcji h_0 nie ma konieczności nakładania wielu założeń, poza tym że powinna ona przyjmować wartości nieujemne. Ponadto funkcja ta zależy tylko od jednej zmiennej, mianowicie czasu t.

Niech $\mathbb{1}_k$ będzie kolumnowym wektorem zerowym, który posiada jedynkę na k-tej pozycji. Zauważmy, że ryzyko względne przy wzroście k-tej współrzędnej charakterystyki x_i dla i-tej obserwacji o jednostkę wynosi

$$\frac{h(t|x_i + \mathbb{1}_k)}{h(t|x_i)} = e^{\beta_k},$$

gdzie $x_i = (x_{i1}, \ldots, x_{ik}, \ldots, x_{ip})^T$ oraz $x_i + \mathbb{1}_k = (x_{i1}, \ldots, x_{ik} + 1, \ldots, x_{ip})^T$.

Dodatkowo warto zwrócić uwagę na ogólną postać ryzyka względnego dla modelu (2.1). Mamy

$$\frac{h(t|x_1)}{h(t|x_2)} = e^{\beta^T (x_1 - x_2)},$$

z czego wynika, że ryzyko względne nie zależy od t. Wyjaśnia to pochodzenie nazwy modelu jako modelu proporcjonalnych hazardów, którego głównym celem jest estymacja tejże wartości.

Powyższe modele zakładają, że zmienna losowa T jest typu ciągłego. Warto wspomnieć, że w przypadku, gdy zmienna losowa T ma rozkład dyskretny, model (2.1) został sprowadzony przez Coxa do uogólnionego modelu regresji logistycznej [6].

Ciekawym zagadnieniem przy analizie różnego typu modeli jest estymacja jego współczynników. Poniżej zaprezentujemy procedurę zaproponowaną przez Coxa, która wykorzystuje funkcję cząstkowej wiarogodności. Ponadto metoda ta nie wymaga jakiejkowiek informacji na temat funkcji hazardu bazowego $h_0(t)$, co niewątpliwie jest jej zaletą.

Niech $R(t) = \{i : t_i > t\}$ oznacza zbiór ryzyka, gdzie t_i są obserwowanymi czasami przeżycia. Zbiór ten obejmuje wszystkie obserwacje, które w chwili t nadal są zagrożone wystąpieniem zdarzenia. Ponadto zakładamy, że dla obserwacji (j), która jest związana z j-tą statystyką pozycyjną, zdarzenie następuje w chwili t_j.

Wtedy funkcja cząstkowej wiarogodności, która jest produktem funkcji największej cząstkowej wiarogodności związanych z m-tą obserwacją, przy założeniach rozpatrywanego modelu jest określona wzorem

$$PL = \prod_{m=1}^{k} \frac{e^{\beta^T x_{(m)}}}{\sum_{x_l \in R(t_m)} e^{\beta^T x_l}} = \prod_{i=1}^{n} \left(\frac{e^{\beta^T x_i}}{\sum_{x_l \in R(t_i)} e^{\beta^T x_l}} \right)^{\delta_i},$$

gdzie k jest liczbą obserwacji niecenzurowanych [7]. Ostatecznie estymator największej wiarygodności wektora współczynników β uzyskujemy po zmaksymalizowaniu zlogarytmowanej funkcji PL.

Wykorzystując estymator $\hat{\beta}$ możemy estymować bazową funkcję hazardu skumulowanego $\hat{H}_0$. Jej estymację umożliwia nieparametryczny estymator największej wiarogodności

zaproponowany przez Breslowa [8]. Przy założeniu, że obserwowane czasy wystąpienia zdarzeń nie nakładają się na siebie przyjmuje on postać

$$\hat{H}_0(t|\hat{\beta}) = \sum_{j:t_j \leqslant t} \frac{\delta_j}{\sum_{k \in R(t_j)} e^{\hat{\beta}^T x_k}}. \tag{2.2}$$

Powyższe wyrażenie jest nazywane estymatorem Breslowa. Został on zbadany dokładniej przez Tsiatisa [9] oraz Andersena [10], którzy przyjrzeli się asymptotycznemu rozkładowi estymatora Breslowa.

Estymator Breslowa może być zdefiniowany również dla bazowej funkcji przeżycia. Określony jest on wzorem

$$\hat{S}_0(t|\hat{\beta}) = e^{-\hat{H}_0(t|\hat{\beta})},$$

gdzie $\hat{H}_0(t|\hat{\beta})$ jest postacią estymatora (2.2). Warto dodać, że w literaturze zaproponowano również inne estymatory funkcji przeżycia. Jednym z nich jest estymator Kalbfleischa-Prentice'a.

Założenie proporcjonalnych hazardów w praktyce często nie jest spełnione. Powoduje to utrudnienia przy próbie wykorzystania modelu w przypadku wielu zastosowań praktycznych. Aby lepiej zilustrować ten problem, przedstawimy pewien przykład. Rozpatrzmy przypadek, w którym porównywane są dwie terapie. Jedna z nich jest terapią właściwą, mającą na celu wyleczenie pacjenta, druga natomiast jest terapią kontrolną, w której pacjenci nie są poddawani żadnej specjalistycznej opiece zdrowotnej. Rozróżnienie tychże grup umożliwia zmienna objaśniająca związana z indykatorem terapii właściwej, tzn. przyjmująca wartość 1 w przypadku zastosowania terapii właściwej, natomiast 0 – w przypadku kontrolnej. Załóżmy, że wartości pozostałych zmiennych objaśniających są takie same w obu przypadkach. Oznaczmy wektory charakterystyk dla leczonych i nieleczonych kolejno przez x_1 oraz x_2. Bez utraty ogólności przyjmiemy, że wartość β związana z indykatorem terapii spełnia warunek $\beta < 0$.

Warunek proporcjonalności oznacza, że funkcja hazardu w przypadku zastosowania terapii leczniczej jest zawsze mniejsza niż dla grupy kontrolnej. Zależność ta zachowana jest także dla skumulowanych funkcji hazardu, mianowicie $H(t|x_1) < H(t|x_2)$. Ponadto ze wzoru $S(t) = P(T > t) = e^{-H(t)}$ natychmiast otrzymujemy, że $S(t|x_1) > S(t|x_2)$. Łatwo widać więc, że jeśli dwie krzywe funkcji przeżycia mają jakiś punkt wspólny, to również odpowiadające im funkcje hazardu będą się przecinać. W szczególności oznacza to, że niestety nie będą one do siebie proporcjonalne. Problem ten jest istotny w praktyce, ponieważ często pojawia się w próbach klinicznych w przypadku, gdy pacjenci są poddani leczeniu radykalnemu. W takiej sytuacji na początku ryzyko śmierci danego pacjenta jest wysokie, lecz w okresie późniejszym, o ile pacjent przeżyje, ulega ono zmniejszeniu i prognozy przeżycia stają się bardziej optymistyczne. Wtedy zaszłaby potrzeba użycia modelu nieproporcjonalnych hazardów.

2.2 Model ze współczynnikiem zależnym od czasu

Model proporcjonalnych hazardów Coxa na przestrzeni lat był w różny sposób modyfikowany. Głównym celem takich modyfikacji było uzyskanie możliwości uwzględnienia przypadku, gdy hazard nie jest proporcjonalny. Jednym z takich zabiegów było wprowadzenie do modelu interakcji pomiędzy zmienną a czasem. Innym pomysłem jest dopuszczenie zależności wektora współczynników od czasu. W rezultacie model ten doskonale radzi sobie z sytuacją, gdy hazard nie jest proporcjonalny. W niniejszym podrozdziale opiszemy

drugą z wymienionych modyfikacji.

Wspomniana wcześniej modyfikacja modelu proporcjonalnych hazardów Coxa zakłada, że

$$h(t|x) = e^{\beta(t)^T x} h_0(t), \tag{2.3}$$

gdzie $\beta(t)$ jest funkcją czasu t. Model (2.3) nazywany jest modelem ze współczynnikiem zależnym od czasu. Zauważmy, że zgodnie z oczekiwaniami ryzyko względne nie jest stałe, ponieważ zależy od czasu i przy wzroście k-tej współrzędnej o jednostkę wynosi

$$\frac{h(t|x_i + \mathbb{1}_k)}{h(t|x_i)} = e^{\beta(t)_k}. \tag{2.4}$$

Ryzyko to jest funkcją nieparametryczną, która zależy od czasu t. Z tego względu z modelem (2.3) wiąże się problem estymacji $\beta(t)$. Przy szacowaniu wartości tego współczynnika wykorzystywane są różne podejścia.

Najprostszym z nich jest metoda estymacji wykorzystująca funkcje schodkowe, lecz procedura ta wymaga podziału osi czasu na różne przedziały. Podział ten z kolei wymaga wyboru odpowiedniej liczby przedziałów, a także odpowiednich ich krańców. Istnieje jednak alternatywna metoda estymacji parametru $\beta(t)$ wykorzystująca tzw. splajny. Poniżej można znaleźć odpowiednią definicję.

Definicja 2.1. Niech $\tau_1 < \tau_2 < \ldots < \tau_K$ będzie zbiorem węzłów tworzących odpowiednie podprzedziały przestrzeni $X \in \mathbb{R}$. Splajnem nazywamy funkcję gładką, która w każdym podprzedziale jest wielomianem stopnia co najwyżej d i posiada ciągłe pochodne rzędu 1, 2, ..., $d-1$ w przedziale, na którym jest określona.

Splajny, które wykorzystuje się w celu estymacji $\beta(t)$ są splajnami kubicznymi. Zgodnie z powyższą definicją są to funkcje, które na przestrzeni $X \in \mathbb{R}$ są dwukrotnie różniczkowalne w sposób ciągły. Ponadto na każdym z podprzedziałów są wielomianami stopnia co najwyżej trzeciego.

Estymacja $\beta(t)$ przy pomocy splajnów kubicznych wiąże się z koniecznością wyboru węzłów, a przede wszystkim ich liczby. Zatem pojawia się tu problem podobny do tego, który miał miejsce w przypadku estymacji za pomocą funkcji schodkowych. Dlatego też w artykule [2] zaproponowano trzecią metodę estymacji, która umożliwia estymację współczynników na podstawie lokalnej zlogarytmowanej funkcji cząstkowej wiarogodności postaci

$$L(\beta, t) = (nh_n)^{-1} \sum_{i=1}^{n} \int_0^{\tau} K\left(\frac{s-t}{h_n}\right) \times \left\{ \beta^T x_i - \log\left(\sum_{j=1}^{n} Y_j(s) e^{\beta^T x_j} \right) \right\} dN_i(s), \tag{2.5}$$

gdzie jądro $K(\cdot)$ jest gęstością rozkładu prawdopodobieństwa o nośniku $[-1, 1]$, średniej 0 i ograniczonej pierwszej pochodnej, $h_n = O(n^{-v})$, gdzie $v > 0$, natomiast τ jest dobrane w taki sposób, aby $P(T_i^* > \tau) > 0$. Funkcja (2.5) jest wypukła ze względu na β. Estymator $\hat{\beta}(t)$ w przypadku, gdy $t \in [h_n, \tau - h_n]$ jest argumentem, który maksymalizuje lokalną zlogarytmowaną funkcję cząstkowej wiarogodności. Jest on rozwiązaniem równania punktowego $U(\beta, t) = 0$ określonego wzorem

$$U(\beta, t) = (nh_n)^{-1/2} \sum_{i=1}^{n} \int_0^{\tau} \{x_i - E(\beta, s)\} \times K\left(\frac{s-t}{h_n}\right) dN_i(s),$$

gdzie

$$E(\beta, t) = S^{(1)}(\beta, t) / S^{(0)}(\beta, t)$$

oraz

$$S^{(r)}(\beta, t) = n^{-1} \sum_{i=1}^{n} Y_i(t) x_i^{\otimes r} e^{\beta^T x_i}, \qquad r = 0, 1, 2.$$

Warto dodać, że estymator $\hat{\beta}(t)$ otrzymany w wyniku zastosowania opisanej metody jest punktowo zgodny dla każdej ustalonej chwili t [2]. Należy jednak zaznaczyć, że zachodzi tu konieczność wyboru szerokości pasma, co w praktyce może okazać się problematyczne.

Niektórzy autorzy zaznaczają również, że często lepszym rozwiązaniem okazuje się być modelowanie ryzyka względnego $h(t|x_i + \mathbb{1}_k)/h(t|x_i)$ za pomocą łatwej w interpretacji funkcji parametrycznej zależnej od czasu.

2.3 Model Yanga-Prentice'a-Diao-Zenga

Model Yanga-Prentice'a-Diao-Zenga, który zostanie opisany w niniejszym podrozdziale, podobnie jak poprzedni nie wymaga założenia proporcjonalności hazardów. Jego nazwa pochodzi od nazwisk czterech statystyków, którzy są jego twórcami [4]. W skrócie jest on często nazywany modelem $YPDZ$ i w ten właśnie sposób będziemy go oznaczać w dalszej części pracy.

Niech x oznacza wektor zmiennych objaśniających. Wtedy model $YPDZ$ zakłada, że

$$\frac{h(t|x)}{h(t|0)} = \frac{e^{\beta_S^T x} e^{\beta_L^T x}}{e^{\beta_S^T x} + (e^{\beta_L^T x} - e^{\beta_S^T x}) S_0(t)}, \tag{2.6}$$

gdzie $S_0(t)$ jest bazową funkcją przeżycia, a β_S oraz β_L są wektorami współczynników modelu. Warto zauważyć, że gdy $\beta_S = \beta_L$ to model Yanga-Prentice'a-Diao-Zenga redukuje się do modelu proporcjonalnych hazardów. Ponadto mamy

$$\frac{h(t = 0|x_i + \mathbb{1}_k)}{h(t = 0|x_i)} = e^{\beta_{S_k}} = \theta_{S_k}$$

oraz

$$\lim_{t \to \infty} \frac{h(t|x_i + \mathbb{1}_k)}{h(t|x_i)} = e^{\beta_{L_k}} = \theta_{L_k}.$$

Wyrażenia oznaczone jako θ_{S_k} oraz θ_{L_k} są interpretowane jako odpowiednio wartość krótkoterminowego oraz długoterminowego ryzyka względnego przy wzroście k-tej współrzędnej x o jednostkę i założeniach modelu (2.6).

Model $YPDZ$ jest bardzo użyteczny w praktyce. W celu ilustracji ponownie posłużymy się przykładem. Podobnie jak poprzednio rozpatrzymy badanych związanych z wartościami danej zmiennej objaśniającej odpowiadającej za indykator terapii właściwej. Podobnie jak poprzednio założmy, że wartości pozostałych ze zmiennych objaśniających są takie same oraz że pacjenci leczeni i nieleczeni są związani odpowiednio z charakterystykami x_1 i x_2. Jeśli wartości krótkoterminowego ryzyka względnego oraz długoterminowego ryzyka względnego są równe jeden, oznacza to, że funkcje hazardu $h(t|x_1)$ oraz $h(t|x_2)$ muszą mieć jakiś punkt wspólny. Z tej obserwacji można wysnuć wniosek, że model $YDPZ$ bardzo dobrze radzi sobie z sytuacją, kiedy funkcje hazardu przecinają się.

Estymacja współczynników modelu $YPDZ$ wymaga wykorzystania wektora postaci $x_i(\cdot)$, który jest związany z i-tą obserwacją i może zależeć od czasu [3]. Warto jednak zauważyć, że metoda ta może również zostać wykorzystana dla modelu (2.6), gdy zmienne nie zależą od czasu. Niech $\bar{x}_i(t)$ będzie zapisem przebiegu zmienności $x(\cdot)$ na odcinku

$[0,t]$. Zakładamy, że $x(\cdot)$ składa się z funkcji ograniczonych i prawostronnie ciągłych. Skumulowany hazard pod warunkiem $\bar{x}(t)$ jest określony wzorem

$$\Lambda(t|\bar{x}_i(t)) = \int_0^t \frac{e^{\beta_S^T x_i(s)} e^{\beta_L^T x_i(s)}}{e^{\beta_S^T x_i(s)} + (e^{\beta_L^T x_i(s)} - e^{\beta_S^T x_i(s)}) S_0(s)} d\Lambda(s),$$

gdzie Λ jest bazową funkcją hazardu skumulowanego. Przy założeniu prawostronnego cenzurowania funkcja wiarogodności jest postaci

$$\prod_{i=1}^{n} \left\{ \frac{e^{\beta_S^T x_i(T_i^*)} e^{\beta_L^T x_i(T_i^*)} \Lambda'(T_i^*)}{e^{\beta_S^T x_i(T_i^*)} + (e^{\beta_L^T x_i(T_i^*)} - e^{\beta_S^T x_i(T_i^*)}) S_0(T_i^*)} \right\}^{\delta_i} e^{-\Lambda(T_i^*|\bar{x}_i(T_i^*))},$$

gdzie $\Lambda'(t)$ jest pochodną $\Lambda(t)$. Estymację współczynników umożliwia maksymalizacja funkcji wiarogodności dla zaobserwowanych danych. Jednakże maksimum to nie istnieje, ponieważ zawsze można wybrać T_i^* takie, że $\Lambda'(T_i^*) = \infty$ dla pewnego T_i^*, $\delta_i = 1$. Skorzystamy z podejścia nieparametrycznego, w którym Λ może być funkcją prawostronnie ciągłą. Dokładniej, zastępujemy $\Lambda'(T_i^*)$ przez $\Lambda\{T_i^*\}$, czyli wysokość skoku $\Lambda(\cdot)$ w T_i^*. Mamy

$$\prod_{i=1}^{n} \left\{ \frac{e^{\beta_S^T x_i(T_i^*)} e^{\beta_L^T x_i(T_i^*)} \Lambda\{T_i^*\}}{e^{\beta_S^T x_i(T_i^*)} + (e^{\beta_L^T x_i(T_i^*)} - e^{\beta_S^T x_i(T_i^*)}) S_0(T_i^*)} \right\}^{\delta_i} e^{-\Lambda(T_i^*|\bar{x}_i(T_i^*))}. \tag{2.7}$$

Maksymalizacja zlogarytmowanego wyrażenia (2.7) metodą Broydena-Fletchera-Goldfarba-Shanno [3] prowadzi do otrzymania estymatorów współczynników modelu oraz bazowej funkcji hazardu skumulowanego.

Znaczącą zaletą modelu jest fakt, że nie wiąże się z nim problem wyboru żadnych parametrów wygładzających. Nie zachodzi również potrzeba podziału osi czasu na części. Ponadto pokazano, że model $YPDZ$ w wielu przypadkach gwarantuje lepsze dopasowanie do danych niż model Coxa. Warto jednak zaznaczyć, że niektóre własności ryzyka względnego związane z modelem $YPDZ$ powodują problemy.

Główny problem, który pojawia się przy wykorzystaniu modelu $YPDZ$ wynika z postaci ryzyka względnego. Dla tego modelu bowiem ryzyko to przy wzroście k-tej współrzędnej x o jednostkę wynosi

$$\frac{h(t|x_i + \mathbb{1}_k)}{h(t|x_i)} = e^{\beta_{S_k}} e^{\beta_{L_k}} \left[\frac{e^{\beta_S^T x} + (e^{\beta_L^T x} - e^{\beta_S^T x}) S_0(t)}{e^{\beta_S^T (x+\mathbb{1}_k)} + \{e^{\beta_L^T (x+\mathbb{1}_k)} - e^{\beta_S^T (x+\mathbb{1}_k)}\} S_0(t)} \right]. \tag{2.8}$$

Łatwo widać, że powyższe wyrażenie zależy od x poza przypadkami, gdy t jest zerem lub dąży do nieskończoności – jest to spowodowane przyjmowaniem w tych przypadkach przez bazową funkcję przeżycia $S_0(t)$ wartości 1 lub 0 odpowiednio. Innymi słowy, w ogólności wartość ryzyka względnego zależy od początkowej wartości x. Z tego względu w dalszej części pracy zostanie przedstawiony model XS będący modelem alternatywnym i niwelującym tą niepożądaną własność modelu $YPDZ$.

2.4 Model XS wykorzystujący interakcje pomiędzy zmienną a czasem

Z powodu wcześniej opisanej zależności ryzyka względnego od x, która miała miejsce w przypadku podstawowej wersji modelu $YPDZ$, powstała potrzeba stworzenia modelu,

który byłby pozbawiony tej wady. W tym celu do modelu proporcjonalnych hazardów Coxa wprowadzono interakcje pomiędzy zmienną a czasem [4]. Mianowicie nowo dodana interakcja jest postaci $xS_0(t)$, gdzie $S_0(t)$ jest bazową funkcją przeżycia.

W rezultacie otrzymujemy model semiparametryczny

$$h(t|x) = e^{\beta_L^T x + \delta^T x S_0(t)} h_0(t), \tag{2.9}$$

nazywany modelem XS wykorzystującym interakcje pomiędzy zmienną a czasem (ang. *XS covariate-time interaction model*). Zauważmy, że dla tego modelu zachodzi

$$\frac{h(t|x_i + \mathbb{1}_k)}{h(t|x_i)} = e^{\beta_{L_k} + \delta_k S_0(t)}. \tag{2.10}$$

Powyższe ryzyko względne nie zależy od x w przeciwieństwie do ryzyka względnego (2.8), które było związane z modelem $YPDZ$. W konsekwencji zaproponowany tu model jest łatwiejszy w interpretacji niż model $YPDZ$.

Przyjrzyjmy się wartościom zarówno krótkoterminowego jak i długoterminowego ryzyka względnego. Z własności funkcji przeżycia wynika, że gdy $t \to \infty$, to $S_0(t) \to 0$. Otrzymujemy zatem

$$\lim_{t\to\infty} \frac{h(t|x_i + \mathbb{1}_k)}{h(t|x_i)} = e^{\beta_{L_k}}.$$

Wartość $e^{\beta_{L_k}}$ może być interpretowana zatem jako wartość długoterminowego ryzyka względnego, które odpowiada wzroście k-tej współrzędnej x o jednostkę.

Przejdźmy do przypadku, gdy $t = 0$. Wtedy z własności funkcji przeżycia natychmiast otrzymujemy, że $S_0(0) = 1$. W konsekwencji mamy

$$\frac{h(0|x_i + \mathbb{1}_k)}{h(0|x_i)} = e^{\beta_{L_k} + \delta_k}.$$

Z przeprowadzonych powyżej rozważań oraz porównując otrzymane tu wyniki z rezultatami uzyskanymi dla modelu $YPDZ$ mamy, że $e^{\beta_{S_k}} = e^{\beta_{L_k} + \delta_k}$ jest krótkoterminowym ryzykiem względnym przy wzroście k-tej współrzędnej x o jednostkę. Zauważmy, że $\delta_k = \beta_{S_k} - \beta_{L_k}$. Wielkość tę możemy wtedy interpretować jako różnicę pomiędzy krótkoterminowymi a długoterminowymi zlogarytmowanymi ryzykami względnymi.

Warto dodać, że z powodów przedstawionych wyżej model (2.9) często nazywany jest nie tylko modelem XS wykorzystującym interakcje pomiędzy zmienną a czasem, ale także modelem szacującym zarówno krótkoterminowe, jak i długoterminowe ryzyko względne. Co więcej, porównując ryzyko względne (2.10) z ryzykami względnymi, które są funkcjami parametrycznymi oraz ryzykiem względnym (2.4) będącym funkcją nieparametryczną, możemy postrzegać tutaj otrzymane ryzyko względne jako pewnego rodzaju alternatywę. Jest to spowodowane faktem, że ryzyko to jest semiparametryczne, zatem łączy w sobie cechy ryzyka parametrycznego oraz nieparametrycznego.

Metoda estymacji współczynników modelu XS wykorzystuje nieparametryczny estymator największej wiarogodności ($NPMLE$) [4]. Co więcej, idea metody jest podobna jak w przypadku estymacji współczynników modelu $YPDZ$. Dla danych cenzurowanych prawostronnie funkcja wiarogodności określona jest wzorem

$$L(\beta_L, \delta, H_0) = \left\{ \prod_{\delta_i=1} h(t_i|x_i) S(t_i|x_i) \right\} \left\{ \prod_{\delta_i=0} S(t_i|x_i) \right\} = \prod_{\delta_i=1} h(t_i|x_i) \prod_{i=1}^{n} S(t_i|x_i), \tag{2.11}$$

gdzie t_i związane są z obserwowanymi czasami przeżycia. Standardowa metoda uzyskania estymatora $NPMLE$ polega na ograniczeniu $H_0(t)$ do klasy funkcji schodkowych, które są rosnące oraz prawostronnie ciągłe. Funkcje te powinny posiadać skoki w momentach $t_{(1)} < \ldots < t_{(K)}$, które oznaczają niecenzurowane czasy przeżycia. W chwili $t = t_{(k)}$ wysokość skoku wynosi h_k, gdzie $1 \leqslant k \leqslant K$. Fakt, że obserwacja t_i jest niecenzurowana pozwala stwierdzić, że $t_i = t_{(k)}$ dla pewnego k. Wtedy $H_0(t_i) = H_0(t_{(k)}) = h_1 + \ldots + h_k = H_k$. Mamy

$$h(t_i|x_i) = h(t_{(k)}|x_i) = e^{\beta_L^T x_i + \delta^T x_i \exp(-H_k)} h_k$$

oraz

$$H(t_i|x_i) = H(t_{(k)}|x_i) = \sum_{j=1}^{k} h(t_{(j)}|x_i) = \sum_{j=1}^{k} e^{\beta_L^T x_i + \delta^T x_i \exp(-H_j)} h_j. \tag{2.12}$$

Wstawiając wyrażenie (2.12) do wzoru

$$S(t_i|x_i) = S(t_{(k)}|x_i) = \exp\left\{-H(t_{(k)}|x_i)\right\}. \tag{2.13}$$

otrzymujemy wartość funkcji przeżycia. W przypadku, gdy t_i jest cenzurowane, jeśli $t_i < t_{(1)}$, to $S(t_i|x_i) = S(0|x_i) = 1$. Jeśli natomiast $t_{(k)} \leqslant t_i < t_{(k+1)}$ dla pewnego k, to przy założeniu, że $t_{(K+1)} = \infty$ mamy, że $S(t_i|x_i)$ jest dane wzorem (2.13). Po wprowadzeniu parametryzacji $\gamma_1 = \log(h_1), \ldots, \gamma_K = \log(h_K)$ zlogarytmowane wyrażenie (2.11) jest maksymalizowane ze względu na β_L, δ oraz $\gamma_1, \ldots, \gamma_K$ przy użyciu metody Broydena-Fletchera-Goldfarba-Shanno [3], co prowadzi do uzyskania estymatorów $NPMLE$.

Rozdział 3

Modele nieproporcjonalnych hazardów

W niniejszym rozdziale zostaną opisane wybrane modele, które podobnie jak wcześniej opisane, są alternatywami dla modelu proporcjonalnych hazardów Coxa. Ich głównym celem jest estymacja związku pomiędzy czynnikami ryzyka a zmienną odpowiedzi. Są one również podstawowym narzędziem wykorzystywanym do prognozy prawdopodobieństwa przeżycia danej jednostki w określonym czasie.

Główny problem, który pojawia się przy wykorzystywaniu modeli zaprezentowanych w tym rozdziale oraz wcześniej opisywanych, to obecność danych cenzurowanych. Jak wiadomo, dane cenzurowane to dane, które nie są kompletne. Jest to równoznaczne z brakiem informacji na temat czasu przeżycia dla niektórych obserwacji. W niniejszym rozdziale ponownie rozpatrywane będą dane cenzurowane prawostronnie, ponieważ są to dane, które często pojawiają się w praktyce. Warto jednak wspomnieć, że w literaturze często prezentowane są zastosowania tychże modeli dla danych cenzurowanych przedziałowo czy też dotyczących różnych zdarzeń, który to czas do ich wystąpienia jest dla statystyków zmienną głównego zainteresowania. Coraz większą popularnością cieszy się również tzw. analiza powtarzanych pomiarów [13].

Oczywiście jak dla każdego innego typu modeli, także dla tych tu opisanych zachodzi konieczność szacowania wartości współczynników. W tym wypadku umożliwia to prognozę zmiennej odpowiedzi, czyli czasu przeżycia. Dla wektora współczynników opisanych modeli istnieje wiele metod estymacji, choć nie każda z nich jest efektywna. Mimo to wybrane procedury estymacji zostały zaprezentowane, co ma głównie na celu zarysowanie złożoności tej problematyki. Obejmują one podejścia od najbardziej powszechnych, jakim jest maksymalizacja funkcji wiarogodności, do rzadziej spotykanych, wykorzystujących ważoną funkcję log-rank czy też maksymalizację prawdopodobieństwa marginalnego rang. Inną wielkością, która jest często estymowana dla tego typu modeli jest funkcja skumulowanego hazardu bazowego.

3.1 Model proporcjonalnych szans

Model proporcjonalnych szans jest modelem często wykorzystywanym w sytuacji, gdy założenie dotyczące proporcjonalności hazardów nie jest spełnione [14]. Ponadto ma on wiele zalet w przypadku, gdy zmienna odpowiedzi przyjmuje wartości ze zbioru skończonego [15].

Załóżmy, że β jest wektorem współczynników modelu, natomiast x jest wektorem

zmiennych objaśniających. Niech G będzie funkcją, o której niczego szczególnego nie zakładamy poza tym, że jest rosnąca oraz, że $G(0) = 0$ [16]. Ponadto niech $S(t|x)$ oznacza warunkową funkcję przeżycia. Wtedy model proporcjonalnych szans jest postaci

$$\text{logit}(1 - S(t|x)) = \log(G(t)) + \beta^T x, \tag{3.1}$$

gdzie $\text{logit}(x) = \log(x/(1-x))$.

Zauważmy, że po prostych przekształceniach można otrzymać wzór na funkcję przeżycia przy założeniach modelu proporcjonalnych szans

$$S(t|x) = \frac{1}{1 + G(t)\exp(\beta^T x)}.$$

Z kolei oznaczając $G'(t) = g(t)$, otrzymujemy wzór na funkcję hazardu

$$h(t|x) = \frac{g(t)}{\exp(-\beta^T x) + G(t)}.$$

Łatwo widać, że ryzyko względne dla jednostek o charakterystykach x_1 oraz x_2 określone jest wzorem

$$\frac{h(t|x_1)}{h(t|x_2)} = \frac{\exp(-\beta^T x_2) + G(t)}{\exp(-\beta^T x_1) + G(t)}.$$

W przypadku, gdy $t = 0$ powyższe wyrażenie upraszcza się do $\exp(\beta^T(x_1 - x_2))$. Z kolei biorąc $t \to \infty$, otrzymujemy w granicy 1, co wynika z faktu, że funkcja $G(t)$ jest rosnąca. Możliwość otrzymania w granicy 1 jest niewątpliwie zaletą modelu proporcjonalnych szans, którą często wykorzystuje się w praktyce [16].

Dodatkowo przedstawimy problem estymacji wektora współczynników modelu, który jest nieodłącznie związany z modelem proporcjonalnych szans. Temat ten był podejmowany przez wielu badaczy takich jak m. in. Dąbrowska i Doksum (1988), Cheng i inni (1995), Murphy i inni (1997) oraz Shen (1998). W celu estymacji używali oni różnych metod. Dla przykładu, Murphy skorzystał z rozwiązania nieparametrycznego. Mianowicie użył on nieparametrycznego estymatora największej wiarogodności ($NPMLE$). Jego podejście zostało opisane poniżej.

Jak wiadomo, w przypadku danych cenzurowanych nie posiadamy pełnej informacji na temat par $\{T_i, x_i\}_{i=1}^n$. Z tego powodu istnieje konieczność rozpatrywania trójek postaci $\{T_i^*, \delta_i, x_i\}_{i=1}^n$, gdzie $T_i^* = \min(T_i, C_i)$. Niech $\Gamma(t) = \exp(G(t))$ oraz $\gamma(t) = d\Gamma(t)/dt$ [13]. Wtedy przy założeniu prawostronnego cenzurowania danych funkcja wiarogodności jest postaci

$$l_n(\beta, \Gamma) = \sum_{i=1}^{n} \{\delta_i[\log \gamma(T_i^*) - \beta^T x_i] - (1 + \delta_i)\log[\Gamma(T_i^*) + \exp(-\beta^T x_i)]\}.$$

Z przyjętych wcześniej założeń wynika, że funkcja $\Gamma(t)$ jest rosnąca. Dodatkowo przyjmujemy, że jest to funkcja schodkowa, posiadająca skoki w chwilach obserwowanych czasów przeżycia. Wielkość tego skoku oznaczamy przez $\gamma\{t\}$, ponieważ wartość $\gamma(t)$ odpowiada za jego wysokość w chwili t. Wtedy funkcja wiarogodności przyjmuje postać

$$l_n(\beta, \Gamma) = \sum_{i=1}^{n} \{\delta_i[\log \gamma\{T_i^*\} - \beta^T x_i] - (1 + \delta_i)\log[\Gamma(T_i^*) + \exp(-\beta^T x_i)]\}.$$

Udowodniono, że maksimum powyższej funkcji istnieje, lecz jego znalezienie wymaga odpowiedniego algorytmu, wykorzystującego pewnego rodzaju iteracje. Murphy pokazał,

że estymator uzyskany w wyniku wykorzystania tejże metody jest zgodny i asymptotycznie normalny.

Inne podejście zostało zaproponowane przez Pettitta. Dzięki jego metodzie możliwa jest estymacja parametru β przy wykorzystaniu maksymalizacji tzw. prawdopodobieństwa marginalnego rang. Warto dodać, że prawdopodobieństwo to nie może być wyliczone bezpośrednio dla każdej współrzędnej wektora β, co skłoniło do wykorzystania aproksymacji zlogarytmowanej funkcji prawdopodobieństwa marginalnego w punkcie będącym wektorem β o wszystkich współrzędnych równych zero, która bazuje na rozwinięciu Taylora. Estymator zaproponowany przez Pettitta nie jest jednak ani nieobciążony, ani zgodny. Lepsze wyniki można uzyskać przy wykorzystaniu metody Lama i Leunga dzięki zastosowaniu specjalnego rodzaju próbkowania. Estymator uzyskany w wyniku zastosowania metody Lama i Leunga wyróżnia się bliskim zeru obciążeniem, lecz niestety zastosowana przez nich procedura jest dość wymagająca obliczeniowo.

3.2 Modele transformacji

W niniejszym podrozdziale zostanie opisana klasa modeli nazywana modelami transformacji. Pozwala ona na analizę rozkładu czasu przeżycia jako funkcji zmiennych objaśniających [17]. Modele transformacji znajdują zastosowanie głównie w analizie przeżycia [18], a ich szczególnymi przypadkami są model proporcjonalnych hazardów oraz model proporcjonalnych szans.

Niech T będzie ciągłą zmienną losową związaną z czasem przeżycia przyjmującą wartości dodatnie, a x wektorem zmiennych objaśniających. Wtedy liniowy model transformacji jest postaci

$$\Gamma(T) = -\beta^T x + \epsilon, \tag{3.2}$$

gdzie Γ jest nieznaną, rosnącą funkcją monotoniczną, β jest wektorem współczynników modelu, natomiast ϵ jest wyrażeniem odpowiadającym za błąd o znanym i ciągłym rozkładzie, które jest niezależne od czasu cenzurowania oraz wektora zmiennych objaśniających [17]. Co więcej, funkcja Γ często nazywana jest czynnikiem skalującym [20].

Warto dodać, że model (3.2) często zapisuje się w równoważnej postaci

$$g(H(t|x)) = G(T) + \beta^T x, \tag{3.3}$$

gdzie g jest znaną funkcją, która zależy od rozkładu zmiennej ϵ, natomiast $H(t|x)$ oznacza skumulowaną funkcję hazardu przy danej zmiennej x. Zapisanie modelu w postaci (3.3) posiada pewne zalety, ponieważ umożliwia uwzględnienie sytuacji, gdy x zależy od czasu.

Jedna z najbardziej popularnych klas modeli, która dopuszcza zależność x od czasu została zaproponowana przez Zenga oraz Lina w 2006 roku [13]. Wykorzystuje ona skumulowane funkcje hazardu dla zmiennej $\{x(s), s \leqslant t\}$ postaci

$$H(t|x) = G\left(\int_0^t \exp\{\beta^T x(s)\} d\Lambda(s)\right),$$

gdzie G jest funkcją ściśle rosnącą, różniczkowalną w sposób ciągły, natomiast Λ jest nieznaną funkcją rosnącą spełniającą warunek $\Lambda(0) = 0$.

W praktyce najczęściej wykorzystuje się tylko wybrane klasy transformacji. Jedną z nich jest klasa transformacji Boxa-Coxa

$$G(x) = ((1+x)^\rho - 1)/\rho,$$

gdzie $\rho \geqslant 0$. W przypadku, gdy $\rho = 0$ zakłada się, że $G(x) = \log(1 + x)$.

Inną klasą transformacji, którą często wykorzystuje się w zastosowaniach praktycznych jest klasa logarytmiczna

$$G(x) = \log(1 + rx)/r,$$

gdzie $r \geqslant 0$. Dla $r = 0$ zakładamy, że $G(x) = x$. Ponadto przy zastosowaniu wspomnianej tu identycznościowej klasy transformacji otrzymuje się model proporcjonalnych hazardów. Przy wykorzystaniu klasy logarytmicznej można również otrzymać model proporcjonalnych szans.

Niektórzy autorzy zastosowali w stosunku do klasy modeli transformacji podejście bardziej ogólne, które wykorzystuje ważone metody odwróconych prawdopodobieństw. Wadą wspomnianego podejścia jest fakt, że zachodzi potrzeba estymacji rozkładu czasów cenzurowania. Dlatego często dokonuje się estymacji odpowiednich równań. Metody tego typu zostały wykorzystane przez Bagdonaviciusa, Nikulina oraz Chena. Oczywiście klasa modeli transformacji ma pewne zalety, jednakże ich zastosowanie zostało zbadane głównie w sytuacjach, gdy mamy do czynienia z wcześniej wspomnianymi przypadkami szczególnymi. W przeciwnym wypadku współczynniki modelu regresji niestety nie są łatwe w interpretacji z powodu ich związku z czynnikiem skalującym [16].

3.3 Model ryzyka addytywnego

W sytuacji, gdy założenie dotyczące proporcjonalności hazardów nie jest spełnione często korzysta się z modelu ryzyka addytywnego, którego popularność w analizie danych cenzurowanych w ostatnim czasie ciągle rośnie [21]. Jednym z jego podstawowych zastosowań jest estymacja różnicy między funkcjami hazardu. Warto zwrócić uwagę, że stwarza to nowe możliwości szacowania zależności pomiędzy nimi. W modelu proporcjonalnych hazardów Coxa estymowana była bowiem wartość ryzyka względnego.

Niech x będzie wektorem zmiennych objaśniających, natomiast β – wektorem współczynników modelu. Podobnie jak poprzednio będziemy obserwować trójki $\{T_i^*, \delta_i, x_i\}_{i=1}^n$.

Przy tych oznaczeniach model ryzyka addytywnego zakłada, że

$$h(t|x) = h_0(t) + \beta^T x,$$

gdzie $h_0(t)$ jest funkcją hazardu bazowego.

Zarówno model proporcjonalnych hazardów Coxa, jak i model ryzyka addytywnego są wykorzystywane w analizie przeżycia jako narzędzia do badania wpływu różnych czynników ryzyka na wystąpienie np. choroby lub śmierci pacjenta [22]. Ponadto zauważmy, że oba te modele są do siebie w pewien sposób podobne, ponieważ wykorzystują funkcję hazardu bazowego [23]. Jednak można zauważyć również pewne różnice. Jedną z nich jest fakt, że w przypadku modelu ryzyka addytywnego wpływ predyktorów na hazard jest liniowy, co można łatwo wynika z postaci modelu. Takie zjawisko nie miało miejsca dla modelu proporcjonalnych hazardów Coxa, gdzie wpływ ten był multiplikatywny.

W kolejnym kroku zaprezentujemy postać estymatora parametru β oraz skumulowanej funkcji hazardu bazowego. Estymator parametru β zaproponowany przez Lina i Yinga jest postaci

$$\hat{\beta} = \left[\sum_{i=1}^{n} \int_0^\infty Y_i(t)(x_i - \bar{x}(t))^{\otimes 2} dt\right]^{-1} \left[\sum_{i=1}^{n} \int_0^\infty (x_i - \bar{x}(t)) dN_i(t)\right],$$

gdzie $\bar{x}(t) = \sum_{i=1}^{n} Y_i(t)x_i / \sum_{i=1}^{n} Y_i(t)$. Ponadto Lin i Ying dowiedli, że estymator $\hat{\beta}$ parametru β jest zgodny oraz asymptotycznie normalny [24].

Przejdźmy do przedstawienia postaci estymatora funkcji skumulowanego hazardu bazowego. Przypomnijmy, że funkcja ta jest dana wzorem $H_0(t) = \int_0^t h_0(u)du$. Przy założeniach modelu ryzyka addytywnego postać estymatora jest następująca

$$\tilde{H}_0(t) = \int_0^t \frac{\sum_{i=1}^{n}(dN_i(u) - Y_i(u)\hat{\beta}^T x_i du)}{\sum_{j=1}^{n} Y_j(u)}. \tag{3.4}$$

Zauważmy, że estymator (3.4) jest całką, a więc w szczególności nie jest funkcją schodkową. Jednak wprowadzając K-elementowy podział osi czasu postaci $t_1 < \ldots < t_K$ otrzymujemy

$$\tilde{H}_0(t_k) = \sum_{l=1}^{k} \frac{\delta_l d_l}{r_l} - \sum_{l=1}^{k} \hat{\beta}^T \bar{x}(t_l)(t_l - t_{l-1}),$$

gdzie d_l i r_l oznaczają kolejno liczbę zdarzeń oraz liczbę obserwacji narażonych na wystąpienie danego zdarzenia w chwili t_l. Co więcej, dla dowolnych $t_k \leqslant t < t_{k+1}$

$$\tilde{H}_0(t) = \sum_{l=1}^{k} \frac{\delta_l d_l}{r_l} - \sum_{l=1}^{k} \hat{\beta}^T \bar{x}(t_l)(t_l - t_{l-1}) - \hat{\beta}^T \bar{x}(t_{k+1})(t - t_k).$$

Niestety otrzymana w wyniku estymacji funkcja przeżycia postaci $\tilde{S}(t|x) = \exp(-\tilde{H}_0(t) - \hat{\beta}^T xt)$ nie spełnia jednego z niezbędnych założeń, ponieważ nie musi być ona nierosnąca. W tej sytuacji zachodzi konieczność zdefiniowania funkcji przeżycia jako $\hat{S}(t|x) = \min_{s \leqslant t} \tilde{S}(s|x)$. Estymator $\tilde{S}$ ten ma pewne zalety, ponieważ jest asymptotycznie równoważny z estymatorem $\hat{S}$. Ponadto $\sqrt{n}(\hat{S}(\cdot|x) - S(\cdot|x))$ zbiega słabo do procesu gaussowskiego o średniej zero. Estymator $\tilde{S}$ ma też jednak pewną wadę. Mianowicie jest on zdefiniowany w sposób, który często powoduje potrzebę szacowania jego wartości w sposób numeryczny.

Ponadto wykazano, że model ryzyka addytywnego bardzo dobrze radzi sobie z pewnymi typami danych i według badań Breslowa i Daya często okazuje się lepszym rozwiązaniem niż model proporcjonalnych hazardów Coxa. Dotyczy to w szczególności danych cenzurowanych przedziałowo [25].

3.4 Model przyspieszonego czasu awarii

Model przyspieszonego czasu awarii jest kolejną alternatywą dla modelu proporcjonalnych hazardów Coxa [26]. Co więcej, ma on również wiele zalet w porównaniu do modelu ryzyka addytywnego [27]. Model przyspieszonego czasu awarii znajduje liczne zastosowania w praktyce, ponieważ wyniki uzyskane dzięki jego wykorzystaniu są łatwe w interpretacji. Doskonale sprawdza się w dziale przemysłu, w szczególności w analizie niezawodności, która polega na badaniu cech różnych obiektów w celu zweryfikowania czy działają one w sposób poprawny przez określony okres czasu i w danych warunkach eksploatacji [28]. Innym przykładem zastosowania jest analiza czasu oczekiwania pieszych na przejście przez jezdnię [29].

Model przyspieszonego czasu awarii zakłada, że

$$\log(T) = \beta^T x + \epsilon, \tag{3.5}$$

gdzie x jest wektorem zmiennych objaśniających, β jest wektorem współczynników modelu, natomiast ϵ jest wyrażeniem odpowiadającym za błąd o nieznanym rozkładzie. Warto wspomnieć, że model przyspieszonego czasu awarii jest modelem liniowym. Często widnieje on w literaturze jako narzędzie wykorzystywane do analizy danych cenzurowanych prawostronnie [27]. Model przyspieszonego czasu awarii zakłada, że zmienne objaśniające mają multiplikatywny wpływ na czas przeżycia [30]. Co więcej, w modelu (3.5) funkcja hazardu dla zmiennej T dana jest wzorem

$$h(t|x) = h_0^* \left(\exp(-\beta^T x)t\right) \exp(-\beta^T x),$$

gdzie h_0^* jest funkcją hazardu dla zmiennej $\exp(\epsilon)$.

Ważnym zagadnieniem, które jest nieodłącznie związane z modelem przyspieszonego czasu awarii jest wybór odpowiedniego rozkładu zmiennej losowej T. Liczne rozważania na temat rozkładów, które są warte zastosowania w praktyce można znaleźć w publikacjach Balakrishnan i Rao oraz Lee i Wanga [29]. Jednym z przykładowych podejść, które często znajduje uzasadnienie w przypadku próby rozwiązania tego problemu jest porównanie wartości AIC (ang. *Akaike Information Criterion*) oraz BIC (ang. *Bayesian information criterion*) dla modeli z wybranymi rozkładami zmiennej T. Takie rozwiązanie zostało zastosowane przez autorów drugiej z wcześniej wymienionych prac. Ostatecznie wybrany został model z najmniejszą wartością wspomnianych miar dopasowania modelu, ponieważ jest to równoznaczne z potencjalnie dość dobrym prognozowaniem danych, które otrzymamy w przyszłości.

W kolejnym kroku przejdziemy do problemu estymacji współczynników modelu. W wyniku procesu cenzurowania obserwujemy realizacje zmiennych $\{T_i^*, \delta_i, x_i\}_{i=1}^n$. Ponadto definiujemy $e_i(\beta) = \log(T_i^*) - \beta^T x_i$, $N_i(\beta; t) = I(e_i(\beta) \leqslant t, \delta_i = 1)$ i $Y_i(\beta; t) = I(e_i(\beta) \geqslant t)$ dla $i = 1, \ldots, n$ [30].

Przy powyższych oznaczeniach parametry modelu mogą być estymowane z wykorzystaniem ważonej funkcji log-rank

$$U_\phi(\beta) = \sum_{i=1}^{n} \delta_i \phi(\beta; e_i(\beta))[x_i - \bar{x}(\beta; e_i(\beta))]$$

lub

$$U_\phi(\beta) = \sum_{i=1}^{n} \int_{-\infty}^{\infty} \phi(\beta; t)(x_i - \bar{x}(\beta; t))dN_i(\beta; t),$$

gdzie ϕ jest funkcją odpowiadającą za wagę, która może zależeć od danych, $\bar{x}(\beta; t) = S^{(1)}(\beta; t)/S^{(0)}(\beta; t)$, $S^{(0)}(\beta; t) = n^{-1}\sum_{j=1}^n Y_j(\beta, t)$ oraz $S^{(1)}(\beta; t) = n^{-1}\sum_{j=1}^n Y_j(\beta, t)x_j$. Przypadki, gdy $\phi(\beta; t) = 1$ oraz $\phi(\beta; t) = S^{(0)}(\beta; t)$ odpowiadają kolejno statystykom log-rank oraz Gehana. Estymator uzyskany przy pomocy tutaj zaprezentowanej metody cechuje się ciekawymi własnościami. Zostały one opisane poniżej.

Niech β_0 będzie prawdziwą wartością wektora współczynników modelu, natomiast $\hat{\beta}_\phi$ – jego estymatorem, czyli rozwiązaniem wcześniej przedstawionych ważonych funkcji log-rank. Wtedy przy pewnych warunkach regularności można pokazać, że wektor losowy $n^{1/2}(\hat{\beta}_\phi - \beta_0)$ jest asymptotycznie normalny ze średnią zero i macierzą kowariancji postaci

$$\Sigma(\beta_0) = A_\phi^{-1}(\beta_0) B_\phi(\beta_0) A_\phi^{-1}(\beta_0),$$

gdzie

$$A_\phi(\beta_0) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \int_{-\infty}^{\infty} \phi(\beta_0; t)\{x_i - \bar{x}(\beta_0; t)\}^{\otimes 2} \times \{\dot{\lambda}(t)/\lambda(t)\} dN_i(\beta_0; t) \tag{3.6}$$

oraz

$$B_\phi(\beta_0) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \int_{-\infty}^{\infty} \phi^2(\beta_0; t) \times \{x_i - \bar{x}(\beta_0; t)\}^{\otimes 2} dN_i(\beta_0; t).$$

W wyrażeniu (3.6) funkcja $\lambda(\cdot)$ oznacza funkcję hazardu dla wyrażenia związanego z błędem, natomiast $\dot{\lambda} = d\lambda(t)/dt$.

Rozdział 4

Modele dla danych z rozkładów semiciągłych

Jak już wcześniej wspomniano wszystkie modele, które opisujemy w niniejszej pracy powinny być wykorzystywane dla danych cenzurowanych prawostronnie. Oznacza to, że znajdują zastosowanie dla danych, w których zdarzenie głównego zainteresowania, takie jak śmierć pacjenta czy przeszczep, może nastąpić po zakończeniu obserwacji pacjenta. Warto wspomnieć jednak, że nie są to jedyne możliwe ograniczenia, które prowadzą do problemów w analizie czasu przeżycia.

Często zdarza się, że dane są nie tylko niekompletne, lecz także pojawia się w nich duża liczba zer (ang. *zero-inflated data*). Zera te dotyczą wartości zmiennej związanej z czasem przeżycia. Przykładem sytuacji, w której takie dane mogą pojawić się z dodatnim prawdopodobieństwem jest analiza czasu do przeszczepu. Niektórzy z pacjentów bowiem z różnych powodów mogą potrzebować przeszczepu natychmiast. Duża liczba zer również może pojawiać się w danych dotyczących czasu przeżycia pacjentów chorych na COVID-19, ponieważ wiele zakażonych pacjentów umiera zaraz po zachorowaniu. W przypadku takiego typu danych standardowo stosowane modele wymagają modyfikacji. Poniżej przedstawiamy proponowane modyfikacje modeli opisanych w poprzednich rozdziałach, które umożliwiają analizę takiego typu danych.

Zasadnicza modyfikacja dotyczy zmiennej odpowiedzi. Zauważmy, że w niniejszej pracy jest ona ciągła. Jednak w sytuacji, gdy w danych znajduje się duża liczba zer proponujemy wprowadzenie semiciągłej zmiennej odpowiedzi Y. Oznacza to, że przyjmuje ona wartości zero z dodatnim prawdopodobieństwem. Dla wektorów losowych X oraz Z, które mogą posiadać pewne elementy wspólne, a ich wartości są wektorami charakterystyk, przyjmujemy, że rozkład warunkowy zmiennej odpowiedzi Y jest określony wzorem

$$F(y|x,z) = P(Y \leqslant y|X = x, Z = z) = \pi(z) + (1 - \pi(z))F_{Y>0}(y|x), \tag{4.1}$$

gdzie $F_{Y>0}(y|x) = P(Y \leqslant y|Y > 0, X = x)$ jest dystrybuantą rozkładu ciągłego, natomiast $\pi(z) = P(Y = 0|Z = z)$. O prawdopodobieństwie $\pi(z)$ niczego szczególnego nie zakładamy. Zauważmy, że rozkład warunkowy określony wzorem (4.1) jest zgodny z intuicją. Mianowicie rozważając skrajną wartość prawdopodobieństwa $\pi(z)$ widzimy, że dla $\pi(z) = 0$ zgodnie z oczekiwaniami rozkład jest ciągły. Najczęściej jednak zmienna Y jest semiciągła.

Idea proponowanej metody polega na przyjęciu modelu $F_{Y>0}(y|x)$. Mianowicie, aby móc użyć modeli opisanych w niniejszej pracy dla danych, w których występuje zerowy

czas przeżycia, należy skorzystać z zależności opisanej wzorem

$$h_{Y>0}(t|x) = \frac{f_{Y>0}(t|x)}{S_{Y>0}(t|x)},$$

gdzie $S_{Y>0}(t|x) = 1 - F_{Y>0}(t|x)$ oraz $f_{Y>0}(t|x) = \frac{d}{dt}F_{Y>0}(t|x)$. Ponadto definiujemy $H_{Y>0}(t|x) = \int_0^t h_{Y>0}(u|x)du$.

W przedstawionych modelach, tzn. modelu proporcjonalnych hazardów Coxa, jego modyfikacjach oraz modelach nieproporcjonalnych hazardów najczęściej należy zastąpić funkcję hazardu wyrażeniem $h_{Y>0}(t|x)$. Czasem jednak istnieje potrzeba modyfikacji poprzez zastąpienie innych wyrażeń pojawiających się w postaci modelu, np. funkcji przeżycia wyrażeniem $S_{Y>0}(t|x)$ w przypadku modelu proporcjonalnych szans czy też funkcji hazardu skumulowanego funkcją $H_{Y>0}(t|x)$ w przypadku modeli transformacji.

Zauważmy, że po przekształceniu modeli w sposób opisany powyżej pojawia się problem związany z estymacją wektora współczynników. Pozostaje to jednak problemem otwartym, wymagającym dalszych badań.

Rozdział 5

Wybór modelu i jego diagnostyka

5.1 Problem wyboru modelu

Problem wyboru modelu sprowadza się do odpowiedniego sprecyzowania zbioru predyktorów, które powinny się w nim znaleźć. Jest to zagadnienie trudne, lecz jednocześnie niezwykle ważne. W celu oceny jakości dopasowania modelu przyjmuje się wiele różnych kryteriów. Często jednak zdarza się, że w wyniku ich zastosowania otrzymuje się odmienne wnioski. Z tego powodu sformułowano kilka ogólnych zasad dotyczących tej problematyki, które powinno brać się pod uwagę w szczególności, gdy wiele zbudowanych modeli wydaje się sensownie prognozować dane. Przede wszystkim model powinien zawierać jak najmniej zmiennych objaśniających. Ich liczba jednak nie może być zbyt mała, ponieważ wtedy model nie będzie dobrze przewidywał wartości zmiennej zależnej. Zasada wyboru jak najmniejszej liczby predyktorów do modelu wynika z prostej obserwacji. Wystarczy bowiem zauważyć, że im więcej zmiennych zawiera model, tym trudniej go zinterpretować. Należy jednak pamiętać, że wybrany model powinien jednocześnie dobrze tłumaczyć zmienną odpowiedzi. Warto zwracać uwagę, jaki wymiar ma ramka danych, którą mamy zamiar wykorzystać do przeprowadzenia analiz. Sytuacja, gdy liczba kolumn jest zbliżona do liczby obserwacji jest niepokojącym zjawiskiem. Wtedy bowiem pojawia się ryzyko przeuczenia modelu (ang. *overfitting*). W konsekwencji model bardzo dobrze prognozuje dane przy wykorzystaniu których został zbudowany, lecz niestety w przypadku nowych danych nie będzie sprawdzał się najlepiej. Jest to kolejny argument przemawiający za wyborem do modelu odpowiednio małej liczby zmiennych.

Wybrane miary oceny jakości dopasowania modelu

W tym rozdziale zostaną opisane miary, które jak do tej pory są jednymi z najczęściej używanych miar oceny jakości dopasowania modelu. Ponadto są one nieodłącznym elementem procedury budowy modelu. Kryterium informacyjne Akaike AIC (ang. *Akaike Information Criterion*) to miara oceniająca jakość dopasowania modelu przy pomocy dywergencji Kullbacka-Leiblera. Dywergencja ta jest tu miarą rozbieżności pomiędzy rozkładami związanymi z rozważanym modelem M oraz nieznanym modelem, którego parametry są estymowane.

Niech $p(y)$ oznacza gęstość rozkładu obserwowalnej zmiennej losowej przy założeniach nieznanego modelu, natomiast $p_M(y, \beta_M)$ – przy założeniach modelu M z wektorem współczynników β_M. Ponadto niech $\hat{\beta}_M$ będzie estymatorem największej wiarogodności parametru β_M, a y^* prognozowaną wartością y [31]. Przy tych założeniach dla rozważanych

rozkładów dywergencja Kullbacka-Leiblera jest określona wzorem

$$KL(p, p_M(\hat{\beta}_M)) = \mathbb{E}\left[\log \frac{p(y^*)}{p_M(y^*, \hat{\beta}_M)}\right].$$

Powyższa wartość oczekiwana jest liczona względem rozkładu związanego z prawdziwą gęstością $p(\cdot)$. Celem opisywanej tu procedury jest minimalizacja wyrażenia $\mathbb{E}[KL(p, p_M(\hat{\beta}_M)]$ dla zbioru wszystkich potencjalnych modeli. Zauważmy, że jest to równoważne minimalizacji $\mathbb{E}[-\mathbb{E}\log(p_M(y^*, \hat{\beta}_M))]$. Akaike udowodnił, że obciążonym estymatorem wyrażenia $\mathbb{E}[\mathbb{E}\log(p_M(y^*, \hat{\beta}_M))]$ jest $L(\hat{\beta}_M)$, gdzie L oznacza logarytm funkcji wiarogodności. Obciążenie to jednak jest redukowane wraz ze zmniejszaniem się liczby predyktorów w modelu M. Z rozważań tych wynika, że rozsądnym wyborem jest model, dla którego minimalizowana jest wartość wyrażenia

$$AIC = -2L(\hat{\beta}_M) + 2k,$$

gdzie k jest liczbą parametrów modelu. Zauważmy, że przy wyznaczaniu wartości AIC brana jest pod uwagę jakość dopasowania oraz złożoność modelu, za którą nakładana jest odpowiednia kara.

Alternatywą dla kryterium informacyjnego Akaike jest bayesowskie kryterium informacyjne BIC (ang. *Bayesian information criterion*). Kryterium to opiera się na minimalizacji wyrażenia

$$BIC = -2L(\hat{\beta}_M) + k \log n,$$

gdzie n jest liczbą obserwacji. Bayesowskie kryterium informacyjne różni się od poprzedniego jedynie wielkością kary nakładanej za złożoność modelu. W przypadku BIC jest ona znacznie większa, czego skutkiem jest częsty wybór modeli o mniejszej liczbie predyktorów.

Powyższy opis dotyczył modeli parametrycznych. W przypadku modeli rozważanych w niniejszej pracy, czyli modeli semiparametrycznych, nie dysponujemy wartością estymatora największej wiarogodności $\hat{\beta}_M$. Nie korzystamy również z funkcji wiarogodności. Proponowane rozwiązania są różne w zależności od rozważanego modelu. W przypadku modelu proporcjonalnych hazardów Coxa $\hat{\beta}_M$ oznacza estymator największej cząstkowej wiarogodności, natomiast $L(\cdot)$ oznacza funkcję cząstkowej wiarogodności. Z kolei dla modelu proporcjonalnych szans funkcja $L(\cdot)$ utożsamiana jest ze zmodyfikowaną funkcją cząstkowej wiarogodności – pseudoprofilowaną funkcją wiarogodności (ang. *pseudo profile likelihood*) dla parametrów regresji.

Metody krokowe

Jednymi z najbardziej popularnych sposobów wyboru zmiennych do modelu są metody krokowe, które polegają na stopniowym dołączaniu bądź usuwaniu predyktorów. Wyróżniamy trzy rodzaje tychże metod: selekcję postępującą (ang. *forward selection*), eliminację wsteczną (ang. *backward elimination*) oraz regresję krokową (ang. *stepwise regression*). Często wykorzystują one wcześniej opisane miary oceny jakości dopasowania.

W przypadku wykorzystania selekcji postępującej najpierw należy zbudować model bez żadnych zmiennych objaśniających. Model taki zawiera jedynie wyraz wolny. Następnie identyfikuje się zmienną objaśniającą, która po dołączeniu do modelu powoduje największy spadek AIC, po to aby w kolejnym kroku dołączyć ją do modelu. Procedura jest zatrzymywana, gdy na danym jej etapie żadna ze zmiennych objaśniających, które można dodać do modelu nie powoduje zmniejszenia wartości AIC. Eliminacja wsteczna przebiega w podobny sposób z tą rożnicą, że modelem startowym jest model ze wszystkimi

możliwymi zmiennymi objaśniającymi, które następnie są kolejno usuwane.

Alternatywą dla metody selekcji postępującej i eliminacji wstecznej jest regresja krokowa, która różni się od standardowych metod krokowych. Oznacza to, że na każdym etapie procedury wcześniej usunięte zmienne mogą zostać ponownie dodane do modelu, natomiast wcześniej dodane zmienne mogą okazać się już nieistotne i wymagać usunięcia. Zaleca się, aby modelem początkowym w tej sytuacji był model ze wszystkimi predyktorami, natomiast nie jest to konieczne. Modelem początkowym może być również model pusty lub jakikolwiek inny.

Istnieje również wiele innych kryteriów decydujących o dodaniu bądź usunięciu danego predyktora. Jedno z nich wykorzystuje wartość SSE (ang. *sum of squared estimate of errors*) oraz test F dla modeli zagnieżdżonych. Najczęściej jednak korzysta się z wcześniej opisanej wersji metod krokowych wykorzystujących wartość AIC. Warto również wspomnieć, że w przypadku zastosowania selekcji postępującej, eliminacji wstecznej oraz regresji krokowej najczęściej otrzymuje się ten sam model.

Indeks $\mathcal{C}$

Indeks $\mathcal{C}$, zwany też współczynnikiem zgodności $\mathcal{C}$ (ang. *concordance index*) jest jedną z najbardziej popularnych miar dyskryminacji dla modeli, które prognozują czas przeżycia [34]. Co więcej, statystyka $\mathcal{C}$ jest ważnym narzędziem, które jest szczególnie przydatne w badaniu różnych modeli predykcyjnych.

Dla danych niecenzurowanych indeks $\mathcal{C}$ jest łatwy w interpretacji, bowiem jest to względna częstość występowania par zgodnych wśród wszystkich par obserwacji. Para zgodna, to taka para obserwacji w której dla jednostki z krótszym czasem przeżycia prognozuje się większe ryzyko wystąpienia zdarzenia.

Niech t będzie ustaloną chwilą czasu, natomiast x_i – wektorem zmiennych objaśniających związanych z i-tą obserwacją. Wtedy przez $M_n(t, x_i)$ oznaczamy prognozowane przez model ryzyko wystąpienia zdarzenia w odpowiednio dobranej chwili t [35]. Przy tych oznaczeniach indeks $\mathcal{C}$ definiuje się jako

$$\mathcal{C} = P(M_n(t, x_i) > M_n(t, x_j) | T_i < T_j).$$

Łatwo spostrzec, że dla danych cenzurowanych prawostronnie pojawia się znacząca trudność, ponieważ chronologia wystąpienia zdarzenia dla pary związanej z danymi tego typu jest niemożliwa do określenia. W literaturze zaproponowano kilka metod estymacji indeksu $\mathcal{C}$ w przypadku takich danych [35]. Jedna z nich polega na zignorowaniu wszystkich par, w przypadku których wcześniejsza obserwacja jest cenzurowana. Wtedy estymator indeksu $\mathcal{C}$ określony jest wzorem

$$\hat{\mathcal{C}}(t) = \frac{\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} I(M_n(t, x_i) > M_n(t, x_j)) I(T_i^* < T_j^*) N_i(t)}{\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} I(T_i^* < T_j^*) N_i(t)},$$

gdzie m jest liczbą obserwacji w zbiorze walidacyjnym, dla którego indeks jest estymowany.

Niech C, D oznaczają liczbę par zgodnych i niezgodnych, natomiast T_x oznacza liczbę par, dla których prognozowany czas przeżycia osiąga te same wartości. Wtedy indeks $\mathcal{C}$ oraz współczynnik korelacji rang d Somersa są powiązane zależnością postaci

$$\mathcal{D} = 2\mathcal{C} - 1,$$

gdzie

$$\mathcal{D} = \frac{C - D}{C + D + T_x}.$$

5.2 Diagnostyka

Przed zbudowaniem modelu warto zidentyfikować obserwacje nietypowe, tzn. takie, które mogą mieć wpływ na model. Często używanym narzędziem jest w tym przypadku odległość Mahalanobisa, która znajduje zastosowanie w identyfikacji obserwacji odstających.

Odległość Mahalanobisa

Jedną z metod umożliwiających identyfikację obserwacji odstających ze względu na wartość zmiennych objaśniających jest kryterium wykorzystujące odległość Mahalanobisa. W ogólności, dla wektorów $x, y \in \mathbb{R}^n$ odległość ta wyraża się wzorem

$$d_n(x, y) = \sqrt{(x-y)C^{-1}(x-y)^T},$$

gdzie C jest macierzą symetryczną i dodatnio określoną. Niech $\{x_1, x_2, \ldots, x_n\}$ będzie rozpatrywaną próbą [38]. Wtedy odległość Mahalanobisa dla i-tej obserwacji jest dana wzorem

$$MD_i^2 = (x_i - \bar{x})^T \Sigma^{-1} (x_i - \bar{x}),$$

gdzie $\Sigma = \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T$. Progiem, który jest często używany dla zaklasyfikowania obserwacji jako odstającej na poziomie istotności α jest $\chi_p^2(1-\alpha)$, gdzie p oznacza liczbę zmiennych [36].

W kolejnym kroku zostaną opisane reszty wykorzystywane do diagnostyki modelu, która jest nieodłącznym elementem procedury mającej na celu stwierdzenie, jak dobrze zbudowany model jest dopasowany do danych. Poniżej opisano różnego rodzaju rezydua, które mają na celu weryfikację założenia proprocjonalności hazardów.

Reszty Schoenfelda

Reszty Schoenfelda zostały wprowadzone w 1982 roku. Podobnie jak wszystkie pozostałe reszty, które opiszemy poniżej, reszty Schoenfelda dobrze radzą sobie w sytuacji, gdy informacja na temat czasu przeżycia nie jest kompletna, co związane jest z obecnością obserwacji cenzurowanych. Podobnie jak w podrozdziale 2.1 definiujemy zbiór ryzyka R wzorem $R(t) = \{i : t_i > t\}$, gdzie t_i oznacza zaobserwowany czas przeżycia. Wprowadźmy oznaczenie

$$E(x_i | R(t_i)) = \frac{\sum_{x_l \in R(t_i)} x_l e^{\beta^T x_l}}{\sum_{x_l \in R(t_i)} e^{\beta^T x_l}}. \tag{5.1}$$

Wtedy reszty Schoenfelda są określone wzorem

$$r_i(\beta) = x_i - E(x_i | R(t_i)),$$

gdzie i jest takie, że $\delta_i = 1$. W praktyce wartość β jest zastępowana wartością $\hat{\beta}$. W ten sposób otrzymywana jest wartość $\hat{r}_i$. W przypadku, gdy założenie proporcjonalności hazardów jest spełnione to $\mathbb{E}\hat{r}_i \approx 0$. Przechodząc do interpretacji graficznej, na wykresie reszt Schoenfelda nie powinniśmy obserwować żadnych trendów. Co więcej, powinny one w przybliżeniu oscylować symetrycznie wokół zera. Bardzo często rozważa się również skalowane reszty Schoenfelda postaci $\hat{r}_i / \hat{V}_i$ [33], gdzie $\hat{V}_i$ jest estymatorem macierzy wariancji.

Reszty Coxa-Snella

Reszty Coxa-Snella to kolejny typ rezyduów za pomocą których można zweryfikować założenie dotyczące proporcjonalności hazardów. Reszta Coxa-Snella dla i-tej obserwacji określona jest wzorem

$$\hat{e}_i = \hat{H}(t_i) = -\log(\hat{S}(t_i)).$$

Przy założeniach modelu proporcjonalnych hazardów ta sama reszta wynosi

$$\hat{e}_i = \hat{H}_0(t|\hat{\beta})e^{\beta^T x}.$$

gdzie $i = 1, \ldots, n$, a $\hat{H}_0(t|\hat{\beta})$ jest estymatorem Breslowa (2.2) funkcji skumulowanego hazardu bazowego. Jeśli założenie proporcjonalnych hazardów jest spełnione, rezydua Coxa-Snella powinny mieć w przybliżeniu rozkład wykładniczy z parametrem 1.

Reszty Lagakosa

Reszty Lagakosa często nazywane są również resztami martyngałowymi. Tematyka reszt Lagakosa była głównie poruszana w pracy statystyka, od nazwiska którego pochodzi ich nazwa (1981) oraz publikacji Barlowa i Prentice'a (1988). Następnie była kontynuowana w rozprawie Therneau (1990). Lagakos stwierdził, że do weryfikacji założenia dotyczącego proporcjonalności hazardów można użyć reszt postaci

$$\hat{M}_i = \delta_i - \hat{e}_i,$$

co w przypadku modelu proporcjonalnych hazardów Coxa sprowadza się do

$$\hat{M}_i = \delta_i - \hat{H}_0(t|\hat{\beta})e^{\beta^T x}.$$

Przy założeniu proporcjonalnych hazardów zachowanie asymptotyczne reszt Lagakosa powinno spełniać $\mathbb{E}(\hat{M}_i) = 0$ oraz $\mathrm{Cov}(\hat{M}_i, \hat{M}_j) = 0$, gdy $i \neq j$ [32]. W ogólności jednak ich rozkład nie jest symetryczny, co prowadzi do trudności w interpretacji. Ponadto reszty te należą do przedziału $(-\infty, 1]$.

W kolejnej części zostaną opisane reszty, które są wykorzytywane w identyfikacji obserwacji odstających i wpływowych. Analiza modelu pod tym kątem jest niezwykle istotna, bowiem eliminuje możliwość nieadekwatnych prognoz z powodu występowania obserwacji, które przyjmują wartości znacząco różniące się od pozostałych. Powody występowania takich obserwacji są bardzo różne. Jednym z nich jest fakt, że w przyrodzie naturalnie można zaobserwować zjawiska nietypowe. Innym przykładem takiej sytuacji jest przypadek, gdy dane zostały wprowadzone do systemu nieprawidłowo.

Reszty dewiacyjne

Reszty dewiacyjne to rodzaj rezyduów, dzięki którym nie tylko można łatwo sprawdzić założenie proporcjonalności hazardów, ale także zidentyfikować obserwacje odstające. Wykorzystują one reszty Lagakosa, które w praktyce okazały się mieć pewne wady. Jest to konsekwencją dużego współczynnika skośności rozkładu, co z kolei spowodowane jest przyjmowaniem wartości z przedziału $(-\infty, 1]$. Therneau (1990) wskazał, że rozsądnym rozwiązaniem tego problemu jest transformacja tychże reszt w taki sposób, aby miały one rozkład zbliżony do normalnego. W swojej publikacji udowodnił, że dzięki temu zabiegowi

można otrzymać lepszą dokładność prognoz.

Reszty dewiacyjne to reszty dane wzorem

$$d_i = \mathrm{sgn}(\hat{M}_i)[-2\{\hat{M}_i + \delta_i \log(\delta_i - \hat{M}_i)\}]^{1/2}.$$

Rezydua te nie mają już wad reszt Lagakosa. Udowodniono bowiem, że gdy w zbiorze danych znajduje się mniej niż 25% obserwacji cenzurowanych to reszty dewiacyjne mają w przybliżeniu rozkład normalny. Bez względu na liczbę obserwacji cenzurowanych rozkład reszt dewiacyjnych powinien symetrycznie oscylować wokół średniej równej zero, a odchylenie standardowe reszt powinno być równe jeden.

Reszty dewiacyjne są pomocne w identyfikacji obserwacji odstających, które pogarszają jakość dopasowania modelu. Duża jej wartość dla jakiejkolwiek obserwacji w porównaniu do pozostałych jest bowiem przesłanką do klasyfikacji jej jako nietypowej [37].

Reszty punktowe (ang. *score residuals*)

Kolejnym rodzajem reszt, który jest bardzo często używany do identyfikacji obserwacji osiągających wartości znacząco różne od pozostałych elementów próby są rezydua punktowe (ang. *score residuals*). Wykorzystują one funkcję punktową $\sum_{\delta_i=1}(x_i - E(x_i|R(t_i)))$, gdzie $E(x_i|R(t_i))$ jest określone wzorem (5.1). Funkcja punktowa jest funkcją, która może zostać wykorzystana do estymacji współczynników modelu proporcjonalnych hazardów Coxa. W konstrukcji reszt punktowych korzysta się z faktu, że funkcja ta może być zapisana w postaci

$$\sum_{i=1}^{n} \int_0^{\infty} (x_i - E(x_i|R(s)))\, d\hat{M}_i(s),$$

gdzie

$$\hat{M}_i(t) = N_i(t) - \int_0^t I(t_i \geqslant s) e^{\hat{\beta}^T x_i} d\hat{H}_0(s)$$

jest procesem martyngałowym.

Wtedy reszty punktowe dla i-tej obserwacji są określone wzorem

$$\int_0^{\infty} (x_i - E(x_i|R(s)))\, d\hat{M}_i(s).$$

Duże wartości tychże rezyduów implikują klasyfikację danej obserwacji jako wpływowej. Co więcej, reszty punktowe podobnie jak reszty Lagakosa cechują się wartością oczekiwaną równą w przybliżeniu zero. Ponadto powinny być one między sobą nieskorelowane [37].

Rozdział 6

Analiza danych rzeczywistych

W niniejszym rozdziale opiszemy analizy danych rzeczywistych przeprowadzone w pakiecie *R* przy wykorzystaniu wybranych modeli. Dane, z których korzystaliśmy w kolejnych podrozdziałach można znaleźć w odpowiednich bibliotekach, których nazwy podaliśmy przy ich opisie. Dodatkowo przed dopasowaniem modelu wykonaliśmy opis danych, który umożliwia wstępne zapoznanie się z nimi. Opis ten obejmuje również niezbędne przekształcenia takie jak obsługa obserwacji brakujących czy zamiana typu zmiennych.

6.1 Model proporcjonalnych hazardów Coxa

Jako pierwsza zostanie przeprowadzona analiza danych przy wykorzystaniu modelu proporcjonalnych hazardów Coxa. Dane, dla których zostanie zbudowany model można znaleźć pod nazwą `ova` w bibliotece `dynpred`. Zawierają one informacje uzyskane w dwóch badaniach klinicznych, które miały na celu porównananie różnych schematów leczenia złośliwego nowotworu jajnika przy wykorzystaniu chemioterapii skojarzonej. Badania te zostały przeprowadzone w Holandii w latach osiemdziesiątych ubiegłego wieku. Wymiar ramki danych to 358×8. Liczba obserwacji wynosi zatem 358, natomiast liczba zmiennych jest równa 8. Poniżej można znaleźć krótki opis rozważanego zbioru danych.

- *tyears* (typ: *numeric*) – czas w latach, który upłynął do śmierci pacjentki lub ostatnio odbytej kontroli lekarskiej,
- *d* (typ: *numeric*) – indykator cenzury (śmierć=1, cenzura=0),
- *Karn* (typ: *numeric*) – skala Karnofsky'ego, która określa stan pacjenta cierpiącego na chorobę nowotworową; im większa wartość wskaźnika tym kondycja chorego jest lepsza,
- *Broders* (typ: *factor*) – skala Brodersa, która jest jednym ze wskaźników wykorzystywanych do oceny stopnia złośliwości nowotworu; występuje na poziomach od 1, 2,3, 4 oraz *unknown*,
- *FIGO* (typ: *factor*) – stopień zaawansowania nowotworu według *FIGO* (*The International Federation of Gynecology and Obstetrics*); występuje na poziomach *III* oraz *IV*,
- *Ascites* (typ: *factor*) – obecność wodobrzusza; występuje na poziomach *absent*, *present* oraz *unknown*,

- *Diam* (typ: *factor*) – średnica guza; występuje na poziomach *micr.*, *<1cm*, *1-2cm*, *2-5cm*, *>5cm*,
- *id* (typ: *integer*) – ID pacjentki.

Na początku zweryfikujemy, czy w zbiorze danych występują obserwacje brakujące. Zmienne *Ascites* i *Broders* występują na poziomach, wśród których pojawia się kategoria *unknown*. Z tego powodu usunęliśmy wszystkie obserwacje, dla których nie mamy informacji na temat występowania wodobrzusza lub wartości skali Brodersa. Operację tę wykonaliśmy przy pomocy poniższego kodu.

```
index1<-which(ova$Ascites=="unknown")
index2<-which(ova$Broders=="unknown")
indexes<-unique(c(index1,index2))
newdata.ova<-ova[-indexes,]
```

Po usunięciu odpowiednich obserwacji otrzymaliśmy, że aktualna ramka danych ma 265 wierszy. W ostatnim etapie procedury należy usunąć poziomy, dla których w konsekwencji liczba obserwacji wynosi zero, ponieważ prowadzi to do późniejszych problemów przy dopasowywaniu modelu.

```
newdata.ova$Ascites<-droplevels(newdata.ova$Ascites)
newdata.ova$Broders<-droplevels(newdata.ova$Broders)
```

Następnie sprawdzimy, czy zmienne są odpowiedniego typu. W realizacji tego zadania pomocna jest funkcja `str()`. W zbiorze `ova` większość zmiennych jest odpowiedniego typu. Jego zamiana powinna być rozważana tylko dla zmiennych *d* oraz *Karn*, ponieważ naturalnym oczekiwaniem jest, aby były one typu *factor*. Jednak sfaktoryzowana zostanie tylko zmienna *Karn*, ponieważ w przypadku zmiennej *d* taka modyfikacja powoduje problemy przy graficznym prezentowaniu danych oraz budowie modelu. Z tych samych powodów w dalszej części pracy nie będziemy faktoryzować zmiennej związanej z indykatorem cenzury.

```
newdata.ova$Karn<-as.factor(newdata.ova$Karn)
```

Dla predyktorów dyskretnych skonstruowaliśmy tabele liczności, które zostały zaprezentowane poniżej.

Tabela 6.1: Tabela liczności dla zmiennej *Karn* ze zbioru danych `ova`

Poziom	Liczba obserwacji
6	14
7	29
8	32
9	85
10	105

Analizując wyniki zawarte w Tabeli 6.1 można stwierdzić, że zmienna *Karn* jest zmienną monotoniczną, bowiem wraz ze wzrostem wartości zmiennej wzrasta również częstość jej

występowania. W ten sposób otrzymujemy, że w rozważanym zbiorze danych znajduje się zdecydowanie więcej pacjentek o bardzo dobrym stanie zdrowia według skali Karnofsky'ego.

Tabela 6.2: Tabela liczności dla zmiennej *Broders* ze zbioru danych `ova`

Poziom	Liczba obserwacji
1	36
2	71
3	112
4	46

W Tabeli 6.2 przedstawiliśmy liczności obserwacji dla zmiennej *Broders*, która odpowiada za stopień zaawansowania nowotworu. Można zauważyć, że niewiele pacjentek związanych jest ze skrajnymi wartościami tejże zmiennej. Co więcej, najmniej z nich jest w początkowym stadium rozwoju choroby. Zdecydowanie najwięcej pacjentek cierpi na nowotwór w trzecim stopniu zaawansowania.

Tabela 6.3: Tabela liczności dla zmiennej *FIGO* ze zbioru danych `ova`

Poziom	Liczba obserwacji
III	201
IV	64

Podobnie jak w przypadku zmiennej *Broders*, w sytuacji, gdy rozważamy stopień zaawansowania nowotworu, lecz tym razem związany z wartościami zmiennej *FIGO*, na podstawie wyników zawartych w Tabeli 6.3 otrzymujemy, że najwięcej pacjentek znajduje się w III stadium zaawansowania choroby.

Tabela 6.4: Tabela liczności dla zmiennej *Ascites* ze zbioru danych `ova`

Poziom	Liczba obserwacji
absent	84
present	181

W Tabeli 6.4 przedstawiliśmy tabelę liczności dla zmiennej *Ascites*. Na podstawie otrzymanych wyników łatwo widać, że u frakcji pacjentek równej 0.68%, czyli większości z nich, zaobserwowano wodobrzusze.

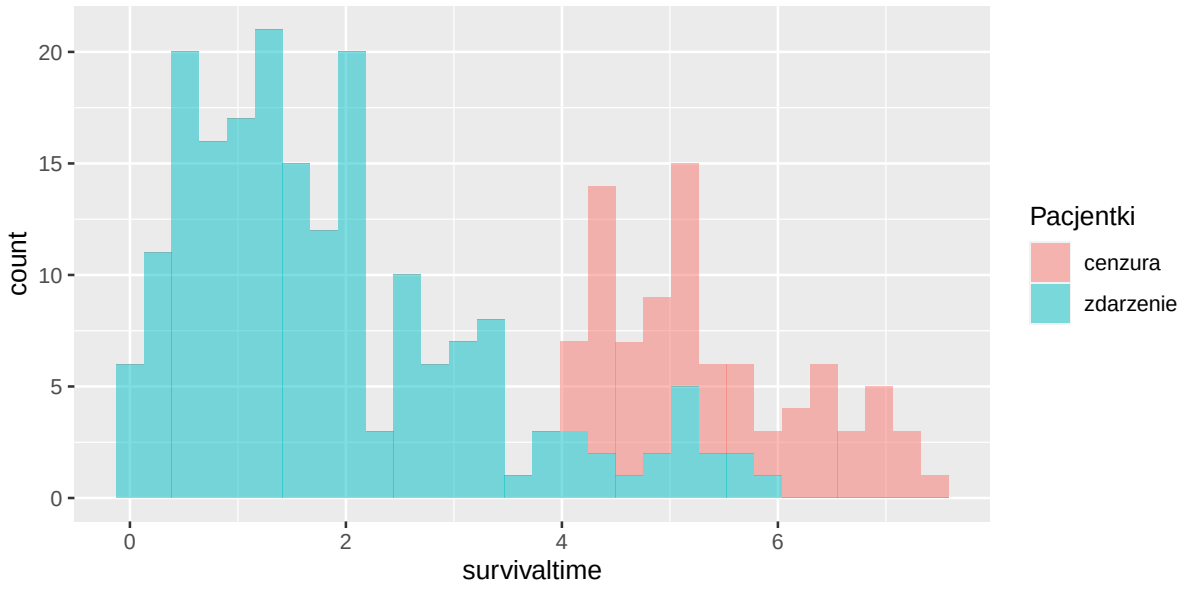
Tabela 6.5: Tabela liczności dla zmiennej *Diam* ze zbioru danych `ova`

Poziom	Liczba obserwacji
micr.	27
<1cm	53
1-2cm	36
2-5cm	50
>5cm	99

Ostatnią rozważaną zmienną kategoryczną, dla której została przedstawiona tabela liczności jest zmienna *Diam*. Wyniki zaprezentowane w Tabeli 6.5 ponownie skłaniają

do wniosku, że liczba pacjentek, dla których ma miejsce zauważalny postęp choroby jest większa. Dla zmiennej *Diam* bowiem obserwuje się najwięcej pacjentek z rozmiarem guza przekraczającym średnicę $5cm$.

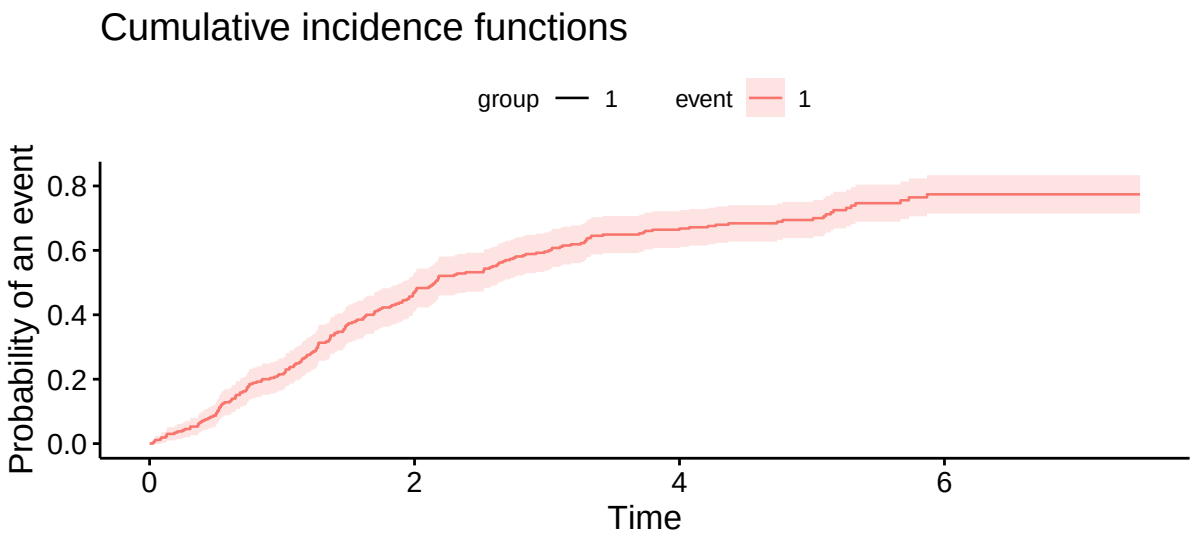
Po przygotowaniu danych i zapoznaniu się z nimi można przejść do ich wstępnej wizualizacji. W kroku początkowym na histogramie przedstawimy liczności dla wszystkich obserwacji z kryterium podziału wyznaczonym przy pomocy zmiennej *d*.



Rysunek 6.1: Histogram czasu przeżycia dla zbioru danych `ova`

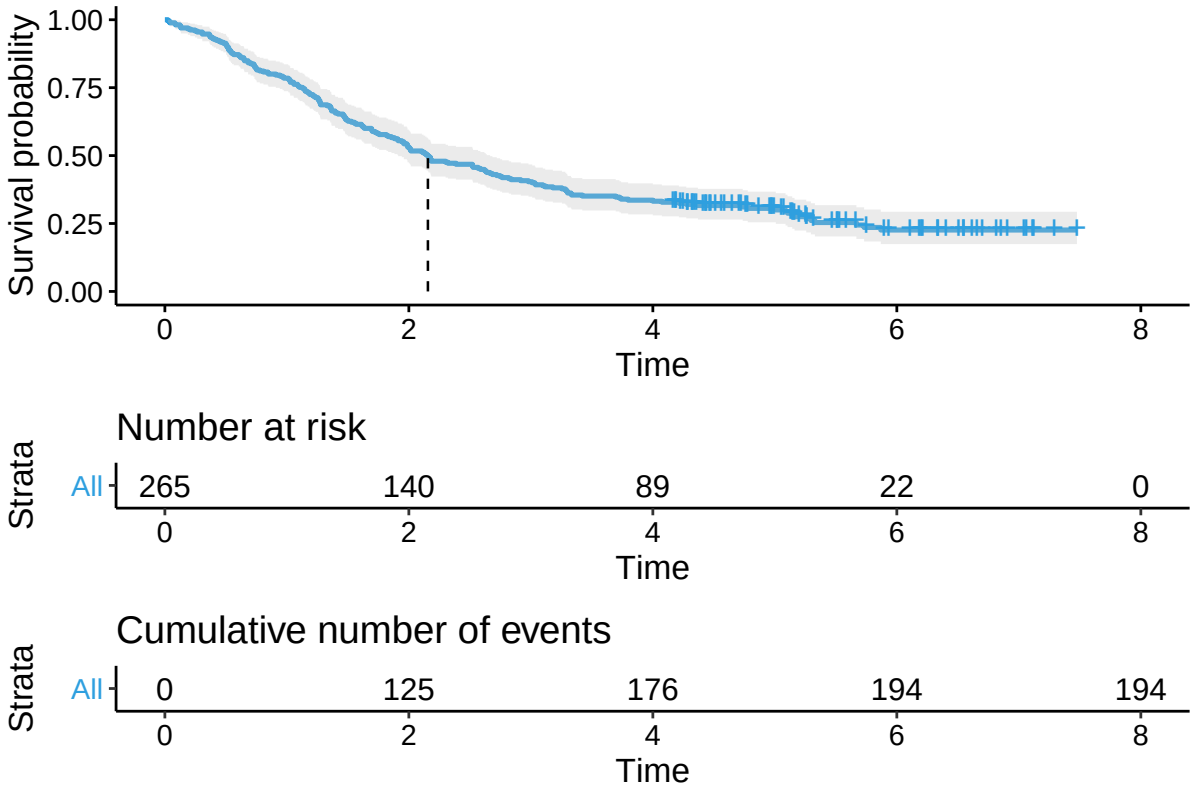
Jak widać na Rysunku 6.1 większa liczba pacjentek związana jest z danymi, dla których zaszło zdarzenie. Ich liczba wynosi dokładnie 194, co stanowi w przybliżeniu 73.21% wszystkich obserwacji. Widać więc, że dane cenzurowane stanowią zdecydowanie mniejszą frakcję obserwacji.

Dodatkowo zbadamy skumulowaną częstość występowania zdarzenia (ang. *cumulative incidence*). Czasem jest ona nazywana także proporcją częstości występowania. Jest to prawdopodobieństwo wystąpienia danego zdarzenia przed danym momentem czasowym. Przypomnijmy, że w przypadku rozważanych tu danych zdarzeniem tym jest śmierć pacjentki z powodu złośliwego nowotworu jajnika bądź wizyta u lekarza. Otrzymane wyniki przedstawiliśmy na Rysunku 6.2.



Rysunek 6.2: Skumulowana częstość występowania zdarzenia dla zbioru danych `ova`

Na Rysunku 6.2 widać, że prawdopodobieństwo zgonu pacjentek z powodu złośliwego nowotworu jajnika bądź odbycia kontroli lekarskiej przed danym momentem czasowym na początku dość szybko wzrasta. Warto dodać, że po ponad 6 latach rozważane prawdopodobieństwo wynosi już prawie 0.8. Można zatem zauważyć, że prawdopodobieństwo wystąpienia zdarzenia przed danym momentem czasowym osiąga wysokie wartości.



Rysunek 6.3: Estymowana funkcja przeżycia wraz z medianą czasu przeżycia dla zbioru danych `ova`

Rysunek 6.3 przedstawia estymowaną funkcję przeżycia. Linią przerywaną zaznaczono medianę czasu przeżycia. Jest to argument, dla którego prawdopodobieństwo przeżycia wynosi 1/2. Po około 6 latach od rozpoczęcia obserwacji prawdopodobieństwo przeżycia wynosi już tylko około 0.25. Ponadto można zauważyć, że po 8 latach nie obserwuje się pacjentek zagrożonych wystąpieniem zdarzenia. Po tym czasie wystąpiły wszystkie z nich.

Po wstępnym zapoznaniu się z danymi możemy przejść do budowy modelu proporcjonalnych hazardów Coxa przy wykorzystaniu funkcji `coxph()`. W kroku początkowym zmienne nie będą wybierane przy pomocy żadnego algorytmu; wykorzystane zostaną bowiem wszystkie z nich.

```
cox.full.model<-coxph(Surv(tyears, d) ~Karn+Broders+FIGO+Ascites+Diam,
data = newdata.ova)
```

Na początku przy użyciu funkcji `cox.zph()` zostanie zweryfikowane założenie proporcjonalności hazardów. Funkcja ta wykorzystuje test, który opiera się na skalowanych resztach Schoenfelda. Hipoteza zerowa w tym teście zakłada, że hazard jest proporcjonalny [39]. Wyniki zaprezentowano w Tabeli 6.6.

```
cox.zph(cox.full.model)
```

Tabela 6.6: Wyniki uzyskane w teście weryfikującym proporcjonalność hazardów dla pełnego modelu proporcjonalnych hazardów Coxa

	Wartość statystyki	Liczba stopni swobody	Wartość p
Karn	3.11	4.00	0.54
Broders	1.28	3.00	0.73
FIGO	0.75	1.00	0.39
Ascites	0.22	1.00	0.64
Diam	6.78	4.00	0.15
GLOBAL	11.12	13.00	0.60

Zgodnie z wynikami przedstawionymi w Tabeli 6.6 na poziomie istotności $\alpha = 0.05$ nie mamy podstaw do odrzucenia założenia o proporcjonalności hazardów. Okazało się zatem, że w tym przypadku można przyjąć założenie o proporcjonalności hazardów. Oznacza to, że model jest dobrze dopasowany i możemy przejść do dalszych analiz.

Po utworzeniu modelu przyjrzymy się osiąganym dla niego wartościom za pomocą funkcji `summary()`. Współczynniki modelu i wartości z nimi związane przedstawiliśmy w Tabeli 6.7.

Tabela 6.7: Współczynniki wraz z wybranymi wartościami dla pełnego modelu proporcjonalnych hazardów Coxa

	β	e^{β}	SE_{β}	z	$P(>\|z\|)$
Karn7	0.13821	1.14822	0.35119	0.39356	0.69390
Karn8	-0.54404	0.58040	0.35957	-1.51302	0.13028
Karn9	-0.70640	0.49342	0.32736	-2.15786	0.03094
Karn10	-0.54757	0.57835	0.31865	-1.71840	0.08572
Broders2	0.72001	2.05446	0.27095	2.65736	0.00788
Broders3	0.63443	1.88595	0.25947	2.44515	0.01448
Broders4	0.40471	1.49887	0.29634	1.36571	0.17203
FIGOIV	0.30206	1.35265	0.16943	1.78285	0.07461
Ascitespresent	0.34723	1.41514	0.16758	2.07202	0.03826
Diam<1cm	0.37238	1.45118	0.34153	1.09031	0.27558
Diam1-2cm	0.90523	2.47249	0.34920	2.59225	0.00953
Diam2-5cm	0.99017	2.69168	0.33728	2.93572	0.00333
Diam>5cm	0.95156	2.58976	0.32188	2.95623	0.00311

Przechodząc do analizy wyników widać, że model zbudowany przy wykorzystaniu wszystkich zmiennych nie jest najlepszym rozwiązaniem. Wynika to z faktu, że dla kilku predyktorów wartość p nie jest ostro mniejsza od poziomu istotności $\alpha = 0.05$, co oznacza, że nie ma podstaw do odrzucenia hipotezy o nieistotności odpowiednich współczynników. Co więcej, wartości błędów standardowych dla poszczególnych współczynników są bardzo duże.

Funkcja `summary()` zwraca również wartości p uzyskane w teście opartym na ilorazie wiarogodności, teście Walda oraz teście logrank. W testach tych hipoteza zerowa zakłada, że wszystkie zmienne w modelu są nieistotne. Otrzymane wyniki są kolejno równe $1.058345e-08$, $5.744545e-08$ oraz $6.601684e-09$. Na poziomie istotności $\alpha = 0.05$ należy odrzucić hipotezę zerową o braku związku predyktorów ze zmienną objaśnianą.

W ostatnim kroku przejdziemy do analizy otrzymanej wartości indeksu $\mathcal{C}$.

Tabela 6.8: Wartości dotyczące indeksu $\mathcal{C}$ dla pełnego modelu proporcjonalnych hazardów Coxa

	Wartość
Indeks $\mathcal{C}$	0.67
C	21367.00
D	10434.00
T_x	215.00

W Tabeli 6.8 zaprezentowano wyniki dotyczące wcześniej wspomnianego indeksu. Otrzymano, że liczba par zgodnych wynosi 21367, liczba par niezgodnych jest równa 10434, natomiast liczba obserwacji, dla których czas przeżycia jest identyczny to 215. Wartość indeksu $\mathcal{C}$ wynosi zatem 0.67 i jest ona dość zadowalająca.

Aby poprawić dopasowanie modelu przejdziemy do wyboru zmiennych do modelu za pomocą metod krokowych. Zastosujemy selekcję postępującą, wsteczną oraz regresję krokową. Procedurę tę umożliwia funkcja `step()`. Oczywiście wcześniej zaszła konieczność zbudowania modelu pustego.

```
cox.empty.model <-coxph(Surv(tyears,d) ~ 1, data = newdata.ova)
```

Następnie możemy przejść do głównej procedury. W przypadku selekcji postępującej model początkowy był modelem pustym, natomiast eliminacji wstecznej oraz regresji krokowej – modelem pełnym. Poniżej przedstawiliśmy przykładowe wywołanie dla selekcji postępującej i kryterium AIC.

```
forward.cox.model.AIC<-step(cox.empty.model,scope=list(upper=cox.full.model,
lower=cox.empty.model), direction="forward",k=2)
```

Modele otrzymane za pomocą wszystkich trzech wariantów i kryterium AIC są identyczne, a ich postać to

```
forward.cox.model.AIC$call

## coxph(formula = Surv(tyears, d) ~ Diam + Karn + Ascites + Broders +
##     FIGO, data = newdata.ova)
```

Poniżej przedstawiliśmy analogiczne wywołanie dla kryterium BIC.

```
n<-sum(newdata.ova$d==1)
forward.cox.model.BIC<-step(cox.empty.model,scope=list(upper=cox.full.model,
lower=cox.empty.model),direction="forward",k=log(n))
```

Podobnie jak poprzednio modele uzyskane za pomocą wszystkich trzech wariantów i kryterium BIC są takie same. Ich postać jest widoczna poniżej.

```
forward.cox.model.BIC$call

## coxph(formula = Surv(tyears, d) ~ Diam + FIGO, data = newdata.ova)
```

Jak wspomniano w podrozdziale 5.1, modele wybrane za pomocą kryterium BIC są zazwyczaj mniej skomplikowane. To zjawisko można zaobserwować również w tym przypadku. Do dalszych analiz z powodów przedstawionych na początku poprzedniego rozdziału wykorzystujemy model uzyskany przy użyciu kryterium BIC.

Dla wybranego modelu przedstawimy wartości, które są zwracane przez funkcję `summary()` i dokonamy ich analizy.

Tabela 6.9: Współczynniki wraz z wybranymi wartościami dla wybranego modelu proporcjonalnych hazardów Coxa

	β	e^{β}	SE_{β}	z	$P(>\|z\|)$
Diam<1cm	0.41236	1.51037	0.33903	1.21628	0.22388
Diam1-2cm	0.99703	2.71021	0.34499	2.88999	0.00385
Diam2-5cm	0.96853	2.63406	0.33234	2.91426	0.00357
Diam>5cm	1.17610	3.24170	0.31181	3.77187	0.00016
FIGOIV	0.45875	1.58209	0.16113	2.84715	0.00441

Dzięki wynikom zaprezentowanym w Tabeli 6.9 możemy w przybliżeniu uznać, że po zastosowaniu metody krokowej wyboru zmiennych do modelu, dla każdej ze zmiennych na poziomie 0.05 należy odrzucić hipotezę o nieistotności. Jedynie w przypadku jednego z poziomów zmiennej *Diam* nie ma podstaw do odrzucenia hipotezy zerowej zakładającej, że zmienna nie jest istotna. Należy jednak zaznaczyć, że ograniczamy się do weryfikacji hipotezy dla zmiennych, a nie ich poziomów. Ponadto można zauważyć, że wartości błędów standardowych uległy zmniejszeniu. Co więcej, widać, że przy założeniach modelu największy wpływ na ryzyko śmierci bądź wizyty lekarskiej ma średnica guza przekraczająca $5cm$. Dla pacjentki ze zdiagnozowanym guzem o takiej średnicy ryzyko to wzrasta aż o około 224% w porównaniu do pacjentek z mikroskopijnym rozmiarem guza. Z kolei w przypadku guza o rozmiarze mniejszym niż $1cm$ analogiczna wartość wynosi 51%. Guz o rozmiarze z przedziału $[1cm, 2cm]$ zwiększa ryzyko o 171%, natomiast w przypadku przedziału $[2cm, 5cm]$ – o 163%. Wartość 58% jest równa wzrostowi ryzyka wystąpienia zdarzenia dla pacjentki, której skala złośliwości nowotworu według $FIGO$ jest stopnia czwartego w porównaniu do chorej, której stopień zaawansowania nowotworu jest o stopień niższy. Oczywiście w rozważaniach tych zakładano, że wartości pozostałych ze zmiennych objaśniających są ustalone.

Dla nowo wybranego modelu wartości p uzyskane we wcześniej już rozważanym teście opartym na ilorazie wiarogodności, Walda oraz logrank są równe kolejno $3.220071e-07$, $1.632015e-06$ oraz $5.405802e-07$. Na poziomie istotności $\alpha = 0.05$ nadal należy odrzucić hipotezę zerową o braku związku predyktorów ze zmienną objaśnianą.

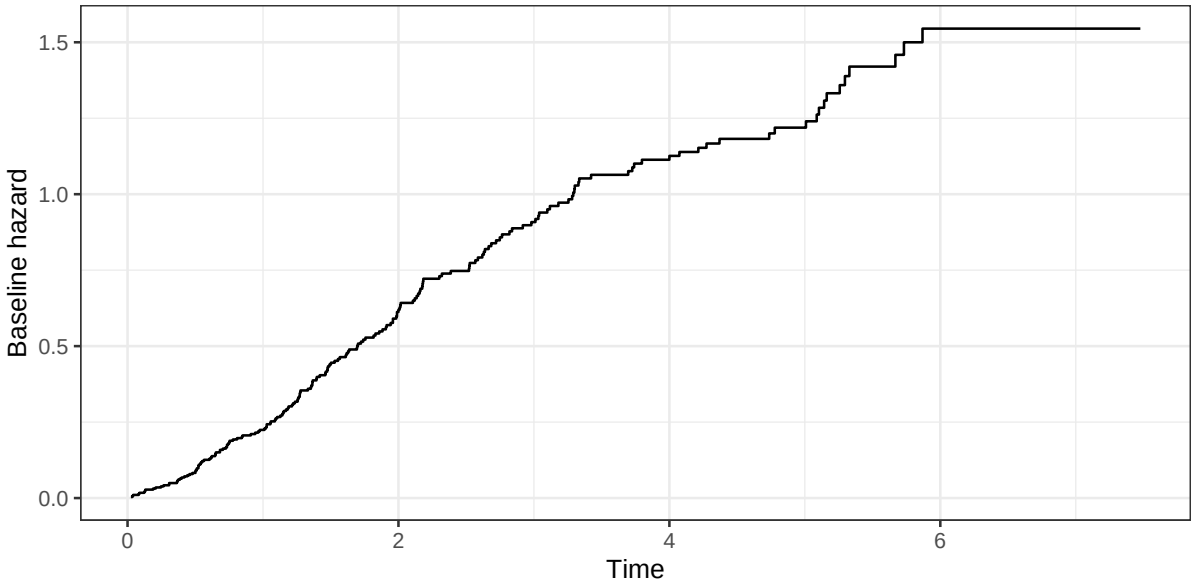
Następnie sprawdzimy jak w porównaniu do poprzedniego modelu zmieniła się estymowana wartość indeksu $\mathcal{C}$. Przypomnijmy, że poprzednio szacowana wartość tego indeksu wynosiła 0.67, a więc była dość zadowalająca.

Tabela 6.10: Wartości dotyczące indeksu $\mathcal{C}$ dla wybranego modelu proporcjonalnych hazardów Coxa

	Wartość
Indeks $\mathcal{C}$	0.64
C	18083.00
D	9241.00
T_x	4692.00

Zgodnie z wynikami przedstawionymi w Tabeli 6.10, dla modelu wybranego przy pomocy kryterium BIC estymowana wartość indeksu $\mathcal{C}$ wynosi 0.64. Widać więc nieznaczne pogorszenie tego wskaźnika porównując otrzymaną wartość z tą uzyskaną wcześniej. Zjawisko to spowodowane jest w głównej mierze wzrostem wartości T_x. Warto zaznaczyć jednak, że spadek wartości rozważanego prawdopodobieństwa jest niewielki, co oznacza, że zbudowany model cechuje się dobrymi własnościami. Warto dodać, że wartość ta nie jest wartością dokładną, a estymowaną z powodu występowania w zbiorze danych licznych obserwacji, dla których dokładny czas przeżycia nie jest znany. Może się zatem zdarzyć, że prawdziwa wartość tego indeksu jest jeszcze wyższa.

Dodatkowo przedstawimy wykres funkcji hazardu bazowego. Rezultaty są widoczne na Rysunku 6.4.



Rysunek 6.4: Funkcja hazardu bazowego dla wybranego modelu proporcjonalnych hazardów Coxa

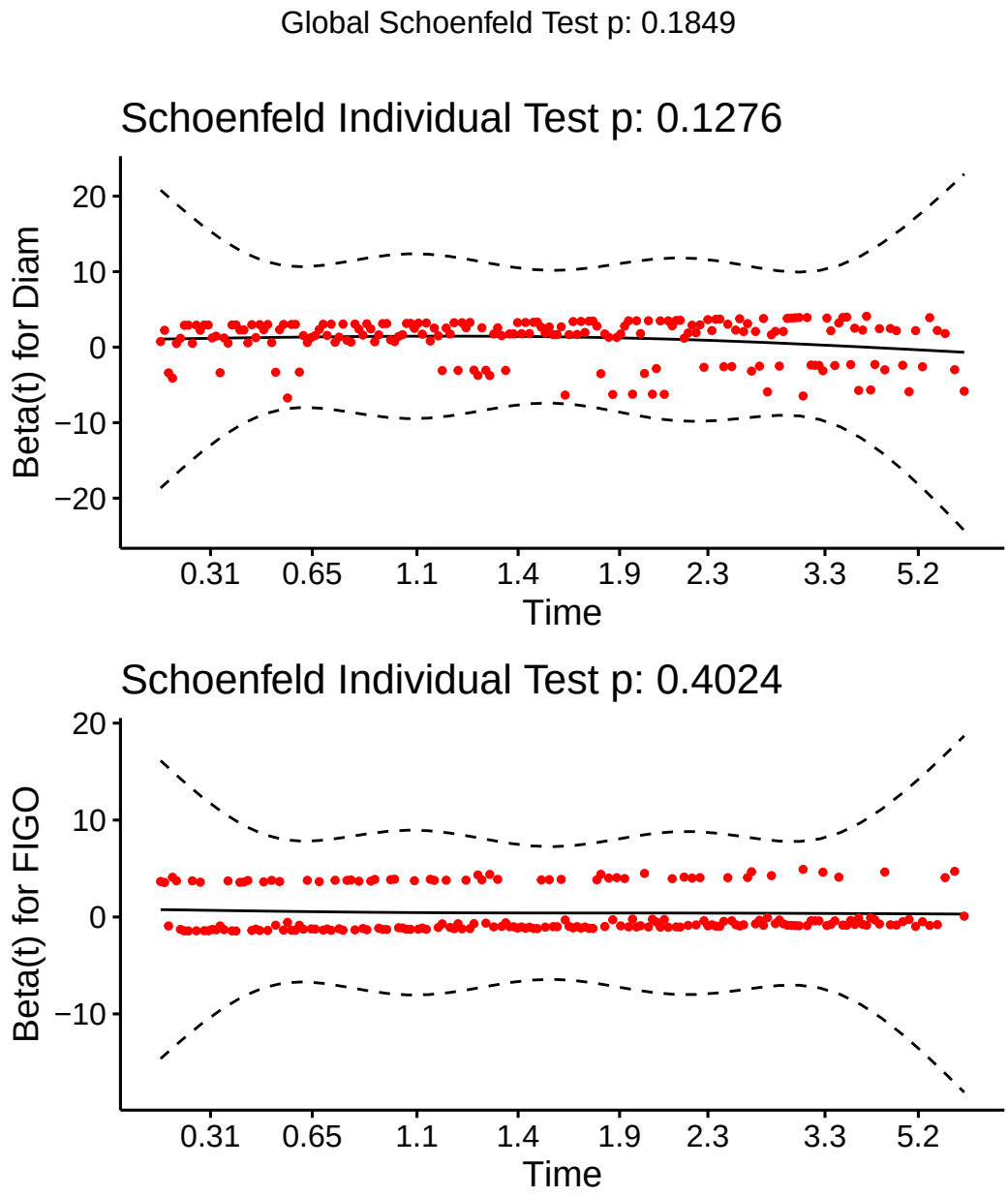
Przypomnijmy, że funkcja hazardu bazowego jest związana z wartościami przyjmowanymi przez funkcję hazardu w przypadku, gdy wartości wszystkich zmiennych objaśniających są równe zero. Zgodnie z oczekiwaniami przyjmuje ona tylko wartości nieujemne.

Po opisie najważniejszych własności modelu przedstawimy wyniki uzyskane w teście weryfikującym założenie proporcjonalności hazardów. Nie mamy tu na celu weryfikacji tejże hipotezy, lecz głównie graficzne przedstawienie otrzymanych wyników. Jednakże dla pełnego zobrazowania sytuacji zamieszczamy wszystkie otrzymane wyniki.

Tabela 6.11: Wyniki uzyskane w teście weryfikującym proporcjonalność hazardów dla wybranego modelu proporcjonalnych hazardów Coxa

	Wartość statystyki	Liczba stopni swobody	Wartość p
Diam	7.16	4.00	0.13
FIGO	0.70	1.00	0.40
GLOBAL	7.52	5.00	0.18

Analizując wyniki zaprezentowane w Tabeli 6.11 ponownie nie mamy podstaw do odrzucenia założenia o proporcjonalności hazardów. Jak wiadomo, model pełny był modelem dobrze dopasowanym dlatego po zastosowaniu metody krokowej wyboru zmiennych do modelu otrzymuje się taki sam wniosek. Dodatkowo wyniki uzyskane w tym teście zostaną zaprezentowane w sposób graficzny. Otrzymane rezultaty są widoczne na Rysunku 6.5.

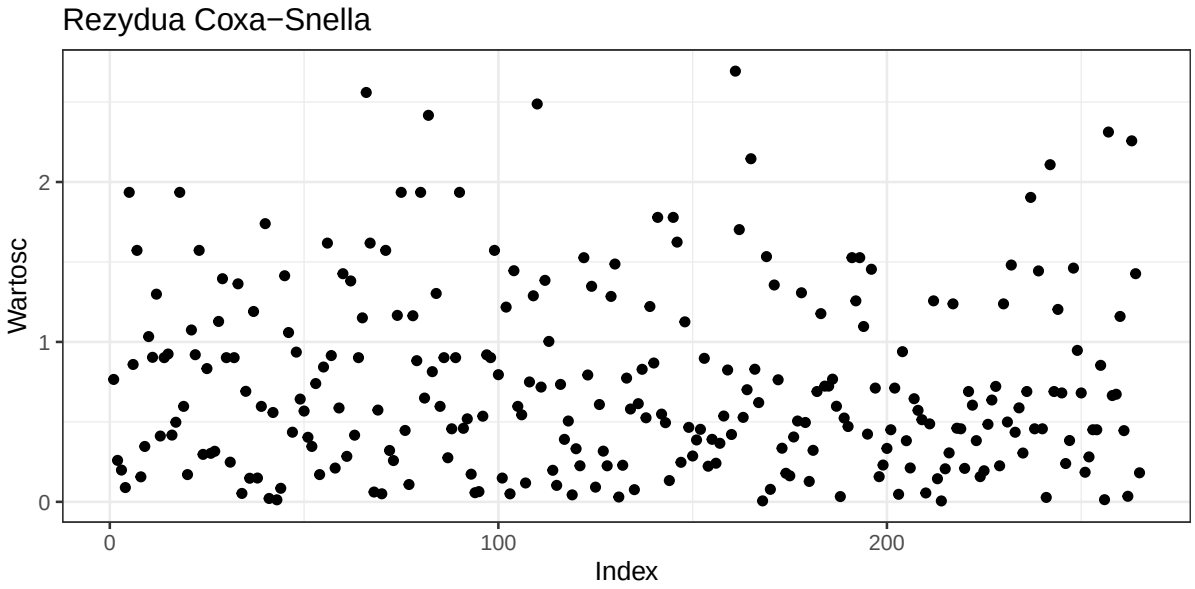


Rysunek 6.5: Rezydua Schoenfelda dla wybranego modelu proporcjonalnych hazardów Coxa

Na Rysunku 6.5 można zobaczyć skalowane reszty Schoenfelda dla nowo wybranego modelu. Liniami przerywanymi zaznaczono przedziały ufności dla dwóch błędów standardowych. Łatwo widać, że rezydua zachowują się zgodnie z oczekiwaniami. Ich rozrzut jest bowiem losowy, a średnia wynosi w przybliżeniu zero.

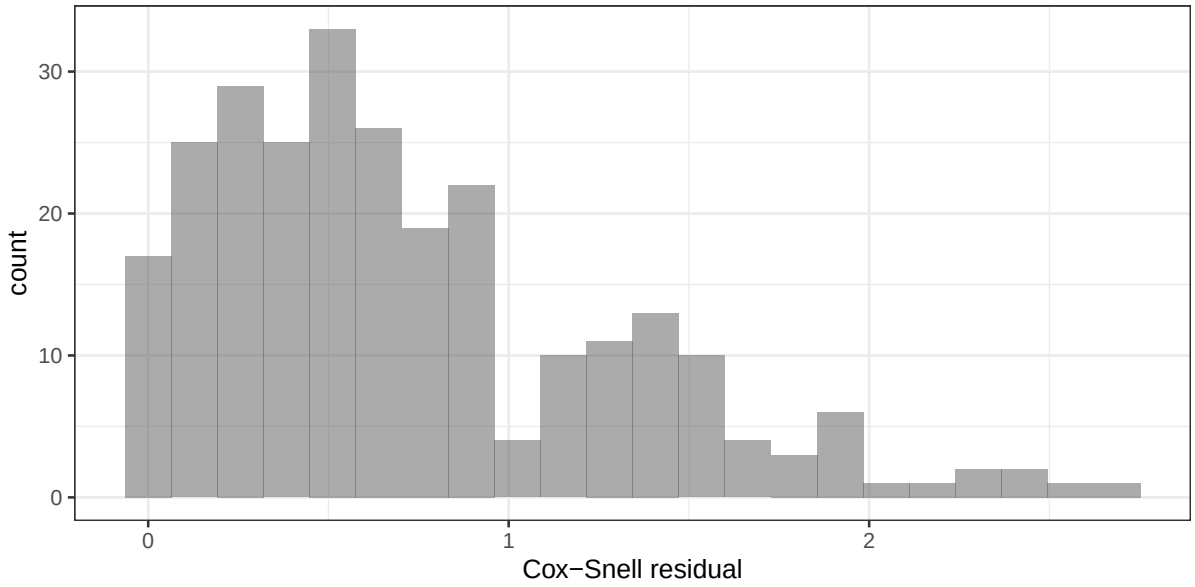
Kolejnym krokiem analizy są reszty Coxa-Snella. W pakiecie *R* nie zaimplementowano żadnej funkcji, która wyliczałaby ich wartość automatycznie. Jednak zgodnie z opisem teoretycznym, który został przestawiony w poprzednim rozdziale skorzystamy z ich związku z resztami Lagakosa.

```
residuals.cox.snell <- newdata.ova$d - martingale.residuals
```



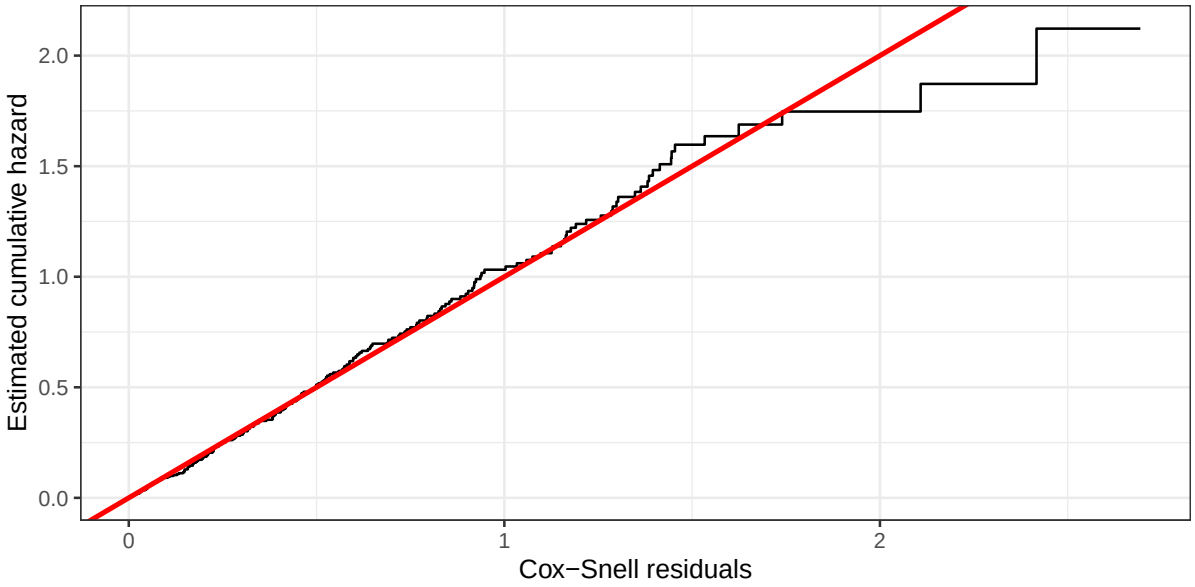
Rysunek 6.6: Rezydua Coxa-Snella dla wybranego modelu proporcjonalnych hazardów Coxa

Na Rysunku 6.6 przedstawiliśmy rezydua Coxa-Snella. Oczekujemy, że będą one pochodzić z rozkładu wykładniczego o średniej 1. We wstępnym kroku weryfikacji tego założenia przyjrzyjmy się rozkładowi tychże reszt.



Rysunek 6.7: Histogram rezyduów Coxa-Snella dla wybranego modelu proporcjonalnych hazardów Coxa

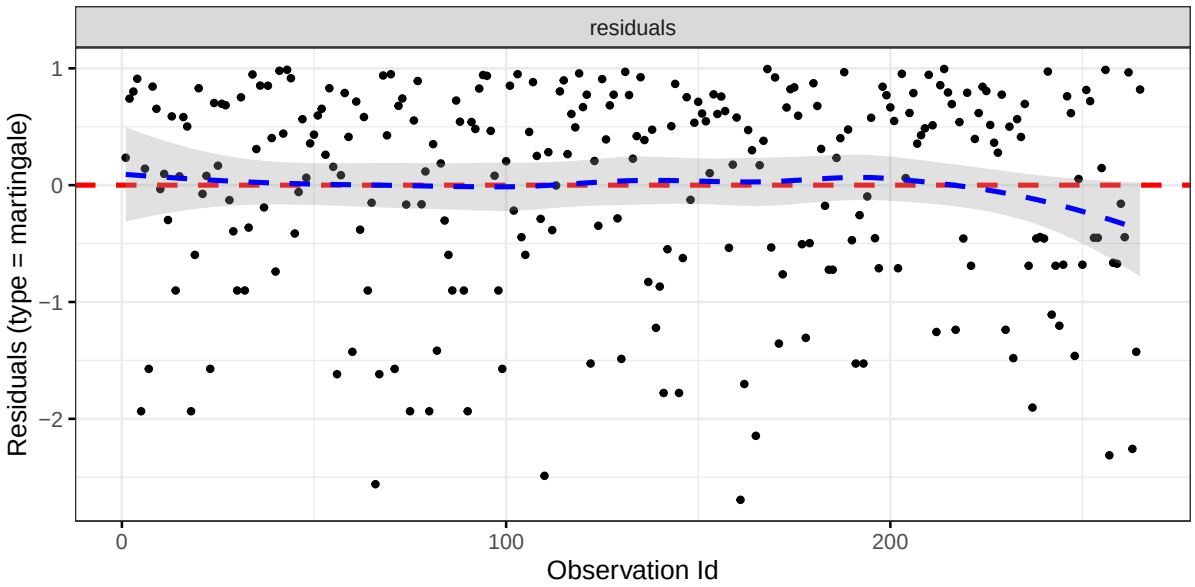
Rozkład przedstawiony na Rysunku 6.7 mimo odchyleń w przybliżeniu przypomina rozkład wykładniczy. Jednak w celu dokładniejszej weryfikacji tego założenia potrzebna jest dalsza analiza. Mianowicie w kolejnym kroku zostanie utworzony wykres skumulowanej funkcji hazardu względem rezyduów Coxa-Snella. Przy założeniu proporcjonalnych hazardów nachylenie otrzymanej krzywej powinno wynosić w przybliżeniu 45°. Wyniki zaprezentowano na Rysunku 6.8.



Rysunek 6.8: Wykres skumulowanego hazardu względem rezyduów Coxa-Snella dla wybranego modelu proporcjonalnych hazardów Coxa

Po raz kolejny możemy stwierdzić, że założenie proporcjonalnych hazardów jest spełnione, ponieważ w przybliżeniu krzywa przedstawiona na Rysunku 6.8 pokrywa się z krzywą czerwoną.

Następnie przejdziemy do analizy reszt Lagakosa.

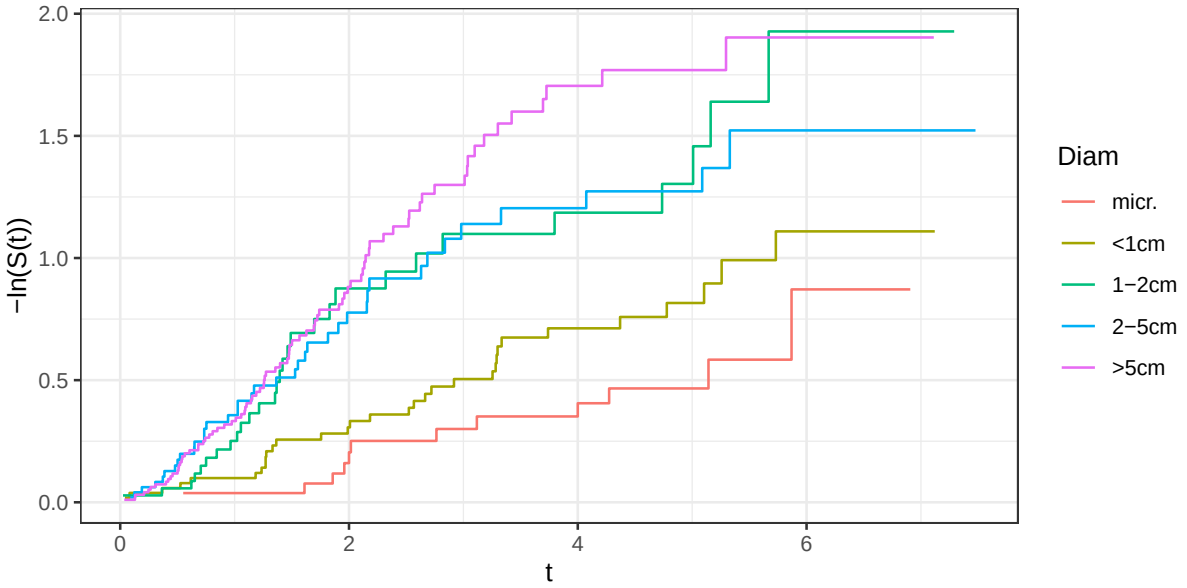


Rysunek 6.9: Rezydua Lagakosa dla wybranego modelu proporcjonalnych hazardów Coxa

Analizując reszty Lagakosa przedstawione na Rysunku 6.9 ponownie należy stwierdzić, że założenie proporcjonalności hazardów będzie spełnione. Na rysunku nie widać bowiem żadnych trendów, a wartość oczekiwana rezyduów jest równa $-2.62917e-16$, więc wynosi w przybliżeniu zero.

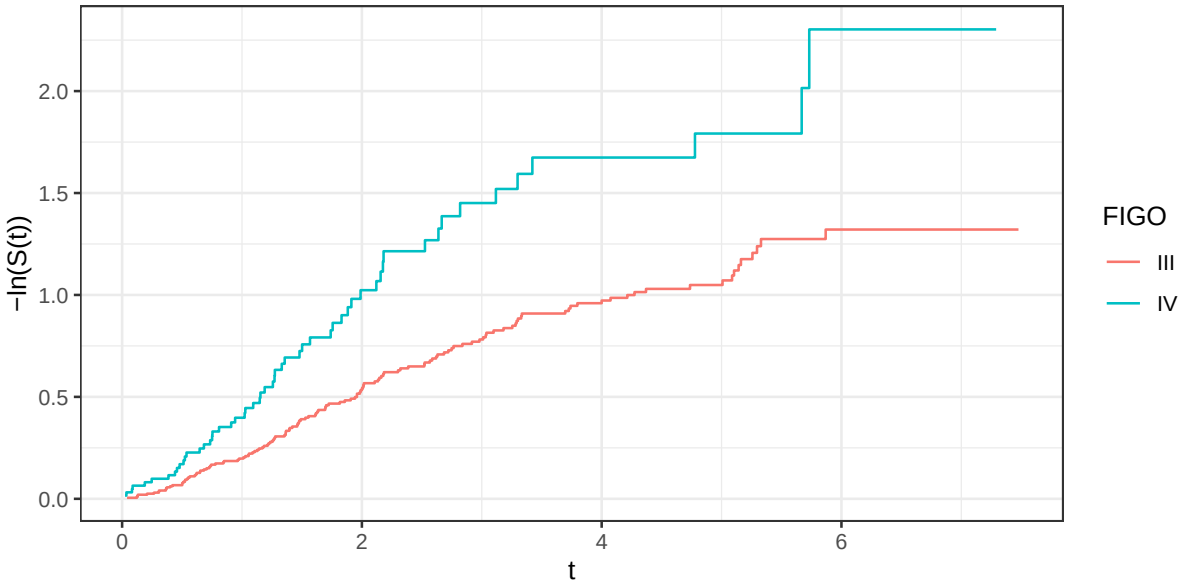
Założenie proporcjonalności hazardów jest również bardzo często weryfikowane przy pomocy analizy wykresu wartości $-\ln S(t)$ względem czasu. Warto zaznaczyć, że oś odciętych

często nie jest tożsama z czasem t, lecz pewną funkcją f zależącą od czasu. Najczęściej wybieraną funkcją, która znajduje zastosowanie w tej sytuacji jest $f(t) = \ln t$. Przy założeniu proporcjonalnych hazardów wykresy wspomnianych zależności otrzymane w podgrupach dla danych zmiennych kategorycznych powinny być w przybliżeniu równoległe. Otrzymane wyniki można zauważyć na Rysunku 6.10 oraz 6.11.



Rysunek 6.10: Wykres zależności $-\ln S(t)$ dla zmiennej *Diam*

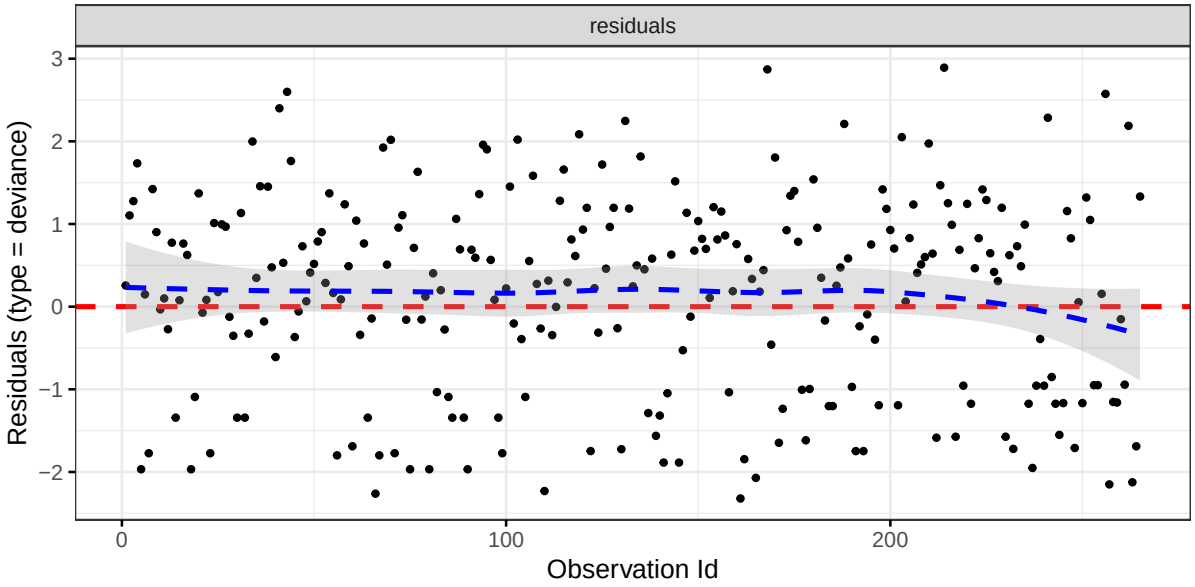
Rysunek 6.10 przedstawia zależność $-\ln S(t)$ względem czasu w podgrupach ze względu na rozmiar guza. Wykres ten nie jest w pełni satysfakcjonujący, lecz należy pamiętać, że jest to jedynie graficzna intepretacja zagadnienia proporcjonalności; wiążący jest dla nas wynik wcześniej już przeprowadzonego testu `cox.zph()`. Co więcej, warto zauważyć, że otrzymany wykres potwierdza, że najmniejszy hazard ma miejsce dla chorych związanych z mikroskopijnym rozmiarem guza.



Rysunek 6.11: Wykres zależności $-\ln S(t)$ dla zmiennej *FIGO*

Rysunek 6.11 ilustruje zależność analogiczną do poprzedniej w podgrupach ze względu na stopień zaawansowania nowotworu według $FIGO$. Proste są do siebie równoległe, co podobnie jak poprzednio pozwala stwierdzić, że hazard spełnia warunek proporcjonalności. Ponadto ponownie zgodnie z oczekiwaniami hazard dla grupy związanej z trzecim stopniem zaawansowania nowotoworu jest mniejszy.

W kolejnym kroku należy zidentyfikować obserwacje, które mogą w wyraźny sposób wpływać na dopasowanie modelu. W tym celu zostaną wykorzystane reszty dewiacyjne oraz punktowe.



Rysunek 6.12: Rezydua dewiacyjne dla wybranego modelu proporcjonalnych hazardów Coxa

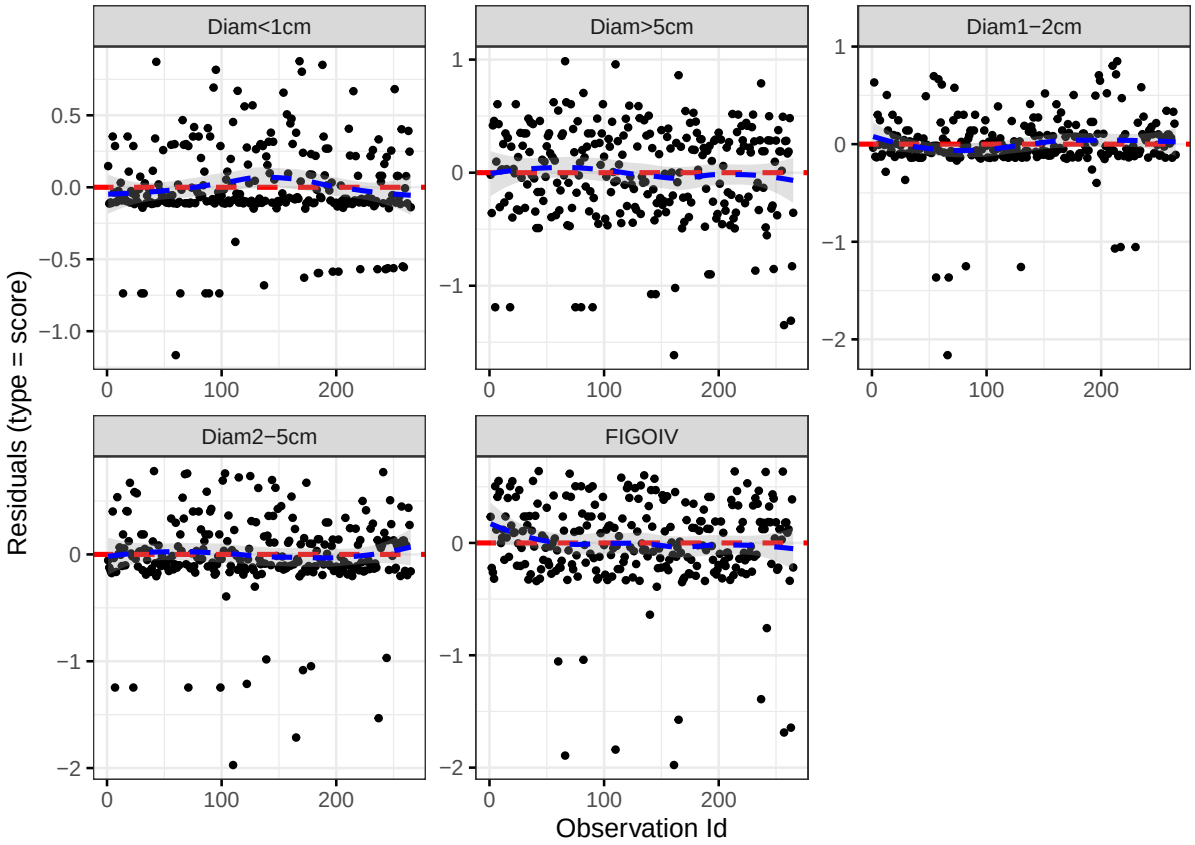
Na Rysunku 6.12 przedstawiliśmy reszty dewiacyjne, które zgodnie z założeniami powinny w przybliżeniu charakteryzować się średnią zero i odchyleniem standardowym jeden. Nie będziemy sprawdzać zgodności rozkładu reszt z rozkładem normalnym, ponieważ dla analizowanego tu zbioru danych frakcja obserwacji cenzurowanych wynosi więcej niż 25%.

```
dev.res<-residuals(forward.cox.model.BIC, type = "deviance")
mean.dev<-mean(dev.res)
sd.dev<-sd(dev.res)
```

Otrzymując wyniki 0.15 oraz 1.2, które odpowiadają średniej oraz odchyleniu standardowemu rezyduów, widać, że potrzebne warunki są spełnione.

Przechodząc do identyfikacji obserwacji odstających widać, że nie mają miejsca żadne odchylenia w wartościach rezyduów i wszystkie z nich należą do przedziału $[-2.5, 3]$. Wnioskujemy więc, że w zbudowanym modelu nie mamy do czynienia z żadnymi obserwacjami odstającymi. Warto dodać, że obserwacje, dla których reszty dewiacyjne są dodatnie, są związane z czasem przeżycia krótszym od oczekiwanego. Analogicznie obserwacje, dla których reszty dewiacyjne są ujemne, związane są z czasem przeżycia dłuższym od oczekiwanego.

Następnie przejdziemy do analizy wartości reszt punktowych. Uzyskane wyniki zostały przedstawione na Rysunku 6.13.



Rysunek 6.13: Rezydua punktowe dla wybranego modelu proporcjonalnych hazardów Coxa

Analizując rezultaty zaprezentowane na Rysunku 6.13 widać, że zgodnie z założeniami wartości oczekiwane reszt dla każdej ze zmiennych są w przybliżeniu równe zero. Dokładne wartości średnich są równe kolejno $1.092898e-10$, $9.310169e-11$, $1.223205e-10$, $2.616519e-10$ oraz $9.505256e-11$.

Dodatkowo sprawdzimy, czy rezydua punktowe dla każdej ze zmiennych są ze sobą skorelowane. Macierz korelacji przedstawiliśmy w Tabeli 6.12.

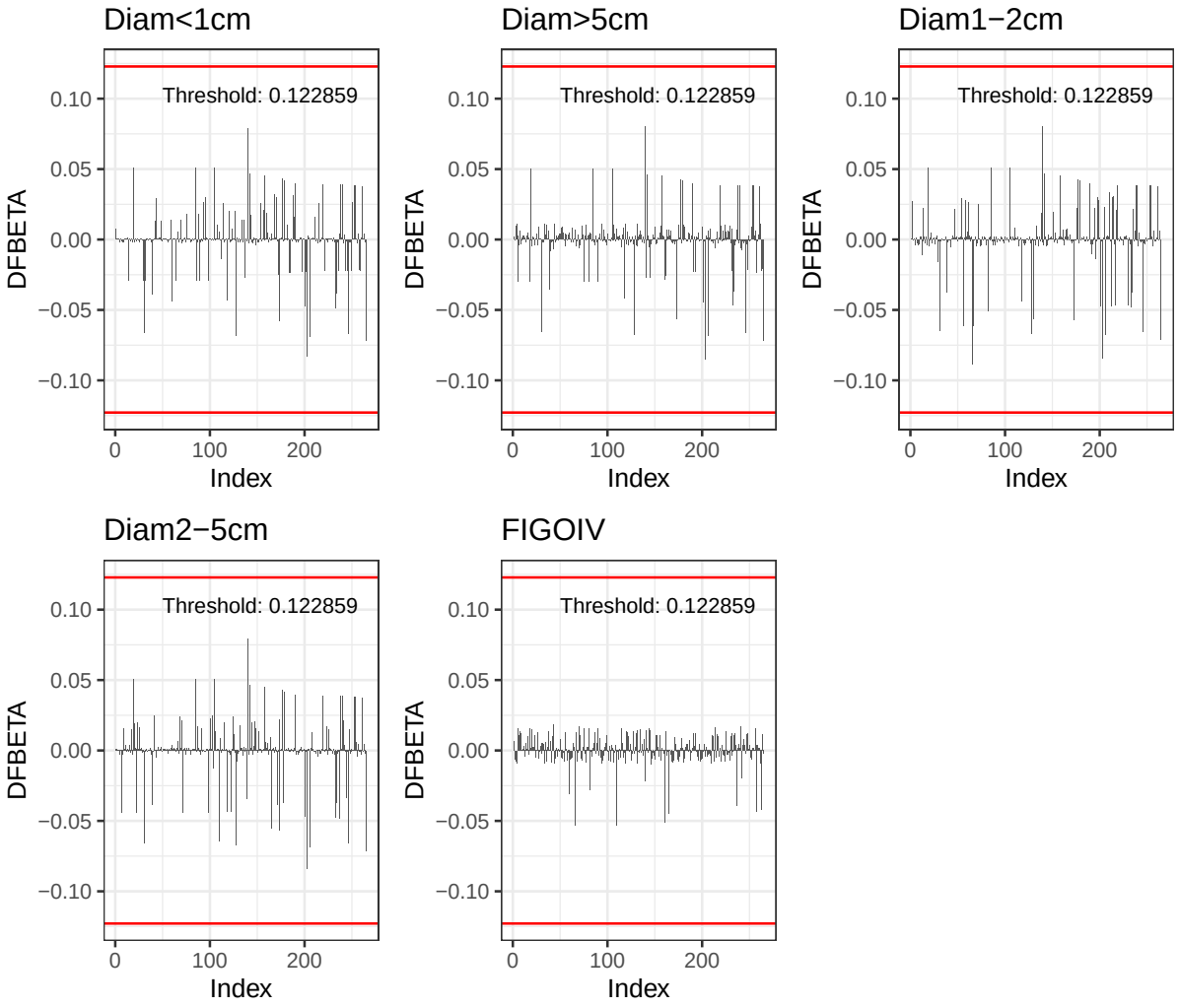
Tabela 6.12: Macierz korelacji reszt punktowych dla wybranego modelu proporcjonalnych hazardów Coxa

	Diam<1cm	Diam1-2cm	Diam2-5cm	Diam>5cm	FIGOIV
Diam<1cm	1.00	-0.17	-0.23	-0.33	-0.12
Diam1-2cm	-0.17	1.00	-0.23	-0.37	0.03
Diam2-5cm	-0.23	-0.23	1.00	-0.48	0.15
Diam>5cm	-0.33	-0.37	-0.48	1.00	-0.02
FIGOIV	-0.12	0.03	0.15	-0.02	1.00

Wyniki zaprezentowane w Tabeli 6.12 pozwalają stwierdzić, że rezydua punktowe dla każdej ze zmiennych nie są ze sobą silnie skorelowane. Współczynniki korelacji co do wartości bezwzględnej nie przekraczają bowiem wartości 0.5, co oznacza, że zmienne są skorelowane w co najwyżej umiarkowanym stopniu [40]. Nie jest to wynik w pełni zadowalający, lecz należy ponownie zaznaczyć, że nie mamy tu do czynienia z silną korelacją.

Przechodząc do identyfikacji obserwacji wpływowych przy pomocy tychże rezyduów łatwo widać, że nie mają miejsca żadne duże odchylenia w wartościach rezyduów punktowych. Oznacza to, że dla zbudowanego modelu nie można zidentyfikować takich obserwacji.

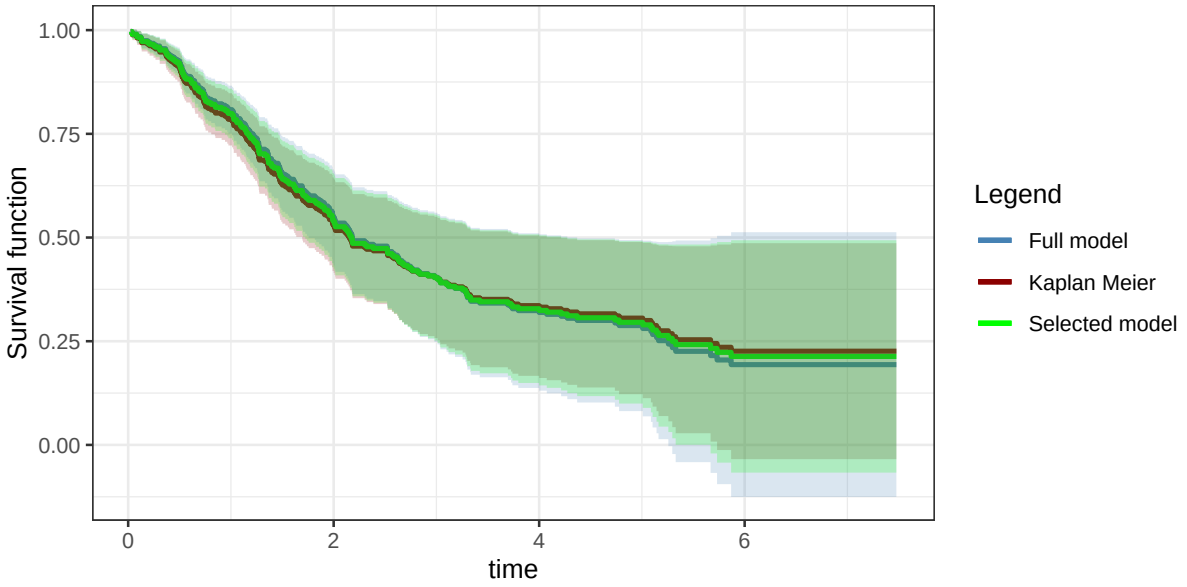
Ostatnim etapem diagnostyki będzie analiza wartości $DFBETA$. Wartości te są często wykorzystywane do identyfikacji obserwacji wpływowych. Są one różnicami pomiędzy współczynnikiem modelu, a tym samym współczynnikiem, lecz po usunięciu i-tej obserwacji. Najczęściej wykorzystywaną wartością progową dla tej procedury jest $2/\sqrt{n}$, gdzie n oznacza liczbę obserwacji. Oczywiście interesująca jest tu wartość bezwzględna $DFBETA$.



Rysunek 6.14: Wartości $DFBETA$ dla wybranego modelu proporcjonalnych hazardów

Z Rysunku 6.14 wynika, że różnica w wartościach współczynników czasem osiąga dość duże wartości, lecz dla żadnej obserwacji nie przekracza progu równego 0.122859. Ponownie nie identyfikujemy zatem żadnych obserwacji wpływowych.

W ostatnim kroku porównano krzywe przeżycia wyestymowane za pomocą estymatora Kaplana-Meiera, modelu ze wszystkimi predyktorami oraz wcześniej wybranego modelu.



Rysunek 6.15: Estymowane funkcje przeżycia dla zbioru danych `ova` i modelu proporcjonalnych hazardów Coxa

Na podstawie wyników przedstawionych na Rysunku 6.15 można stwierdzić, że wszystkie estymowane funkcje przeżycia osiągają podobne wartości, w szczególności w ciągu około 3 lat od rozpoczęcia badań. Po tym okresie czasu różnice w estymowanych prawdopodobieństwach przeżycia są bardziej zauważalne, choć nadal niewielkie. Co więcej, łatwo widać, że w szczególności w późniejszym okresie badań, szacowane prawdopodobieństwa przeżycia osiągają lepsze wartości w przypadku modelu z predyktorami *Diam* oraz *FIGO*, ponieważ estymowana funkcja przeżycia jest bliższa estymatorowi Kaplana-Meiera.

Jak już wcześniej wspominaliśmy w podrozdziale 2.1, założenie proporcjonalności w praktyce często nie jest spełnione. Jednak zdarza się, że dla pewnych danych model proporcjonalnych hazardów Coxa znajduje zastosowanie. Taka sytuacja miała miejsce w przypadku analizy przeprowadzonej w niniejszym podrozdziale. Należy jednak zaznaczyć, że w ogólności założenie proporcjonalności hazardów powoduje problemy, dlatego w kolejnych podrozdziałach przeprowadzimy analizy, które są stosowane w tego typu sytuacjach.

6.2 Model ze współczynnikiem zależnym od czasu

W niniejszym podrozdziale przedstawimy procedurę dopasowania modelu ze współczynnikiem zależnym od czasu. Rozpatrywać będziemy dane `tTRACE` z pakietu `timereg`. Dane pochodzą z badania przeprowadzonego dla 5157 pacjentów, którzy przebyli zawał mięśnia sercowego i byli leczeni w szpitalu Glostrup w Danii w latach 1977 – 1988. Zbiór `tTRACE` jest podzbiorem tychże danych obejmującym 1000 pacjentów oraz 9 zmiennych, które opisano poniżej.

- *id* (typ: *integer*) – ID pacjenta,
- *wmi* (typ: *numeric*) – miara efektywności pompowania krwi przez serce otrzymana w badaniach używających fale ultradźwiękowe; im wyższa wartość tego wskaźnika, tym lepsza wydolność serca,

- *status* (typ: *numeric*) – status pacjenta (śmierć z powodu zawału mięśnia sercowego=9, pacjent żyjący=0, śmierć z innych powodów=7),
- *chf* (typ: *integer*) – indykator klinicznej niewydolności pompy serca,
- *age* (typ: *numeric*) – wiek pacjenta,
- *sex* (typ: *integer*) – płeć pacjenta (kobieta=1, mężczyzna=0),
- *diabetes* (typ: *integer*) – indykator cukrzycy,
- *time* (typ: *numeric*) – czas przeżycia w latach,
- *vf* (typ: *numeric*) – indykator migotania komór.

W zbiorze danych nie zidentyfikowano żadnych obserwacji brakujących. Jednak nie wszystkie zmienne są odpowiedniego typu, ponieważ zmienne *chf*, *sex*, *diabetes* oraz *vf* powinny zostać sfaktoryzowane. Zamianę typu przeprowadziliśmy za pomocą poniższego kodu.

```
tTRACE$chf<-as.factor(tTRACE$chf)
tTRACE$sex<-as.factor(tTRACE$sex)
tTRACE$diabetes<-as.factor(tTRACE$diabetes)
tTRACE$vf<-as.factor(tTRACE$vf)
```

W kolejnym kroku wyliczyliśmy wartości średniej, wariancji, odchylenia standardowego, mediany, pierwszego i trzeciego kwartyla, minimum oraz maksimum dla wszystkich zmiennych, które nie są jakościowe.

Tabela 6.13: Wartości wskaźników dla zmiennych ciągłych ze zbioru danych `tTRACE`

	wmi	*age*
Średnia	1.4075	67.4201
Wariancja	0.1625	130.1282
Odchylenie standardowe	0.4031	11.4074
Mediana	1.4000	68.5570
Pierwszy kwartyl	1.1000	60.0613
Trzeci kwartyl	1.8000	75.6010
Minimum	0.3000	22.9440
Maksimum	2.2000	96.3320

Z wartości wskaźników zawartych w Tabeli 6.13 wynika, że średnia wieku badanych wynosiła około 67 lat. Ponadto największe zróżnicowanie w zebranych wynikach obserwuje się dla zmiennej związanej z wiekiem. Wartości mediany są zbliżone do wartości średnich. Dla 25% pacjentów zarejestrowano wartość zmiennej *wmi* mniejszą bądź równą 1.1. Ta sama grupa obserwacji była związana z wiekiem mniejszym bądź równym około 60.06 lat. Podobnie 75% pacjentów było związanych z wartością zmiennej *wmi* mniejszą bądź równą 1.8 oraz wiekiem mniejszym niż około 75.6 lat. Wartości zmiennej *wmi* mieszczą się w przedziale $[0.3, 2.2]$, a wiek badanych pacjentów wynosił od prawie 23 lat do niemal 100 lat.

Kolejnym krokiem we wstępnym zapoznaniu się z danymi są tabele liczności dla predyktorów jakościowych.

Tabela 6.14: Tabela liczności dla zmiennej *chf* ze zbioru danych `tTRACE`

Poziom	Liczba obserwacji
0	468
1	532

Wyniki przedstawione w Tabeli 6.14 pozwalają stwierdzić, że 532 pacjentów dotyczy kliniczna niewydolność pompy serca. Warto zauważyć, że jest to ponad połowa wszystkich obserwacji.

Tabela 6.15: Tabela liczności dla zmiennej *sex* ze zbioru danych `tTRACE`

Poziom	Liczba obserwacji
0	310
1	690

Z Tabeli 6.15 wynika, że większość obserwacji w rozpatrywanym zbiorze danych stanowią kobiety. Ich frakcja jest równa 69%.

Tabela 6.16: Tabela liczności dla zmiennej *diabetes* ze zbioru danych `tTRACE`

Poziom	Liczba obserwacji
0	906
1	94

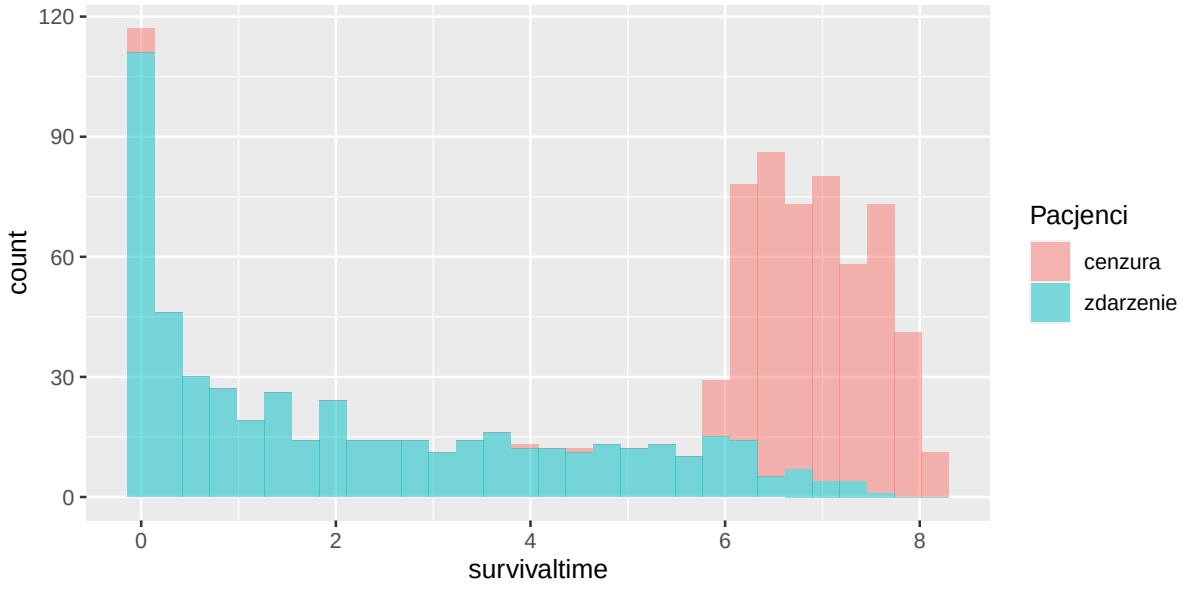
Analizując wyniki przedstawione w Tabeli 6.16 możemy stwierdzić, że mniej niż 10% pacjentów cierpi na cukrzycę. Aż u 906 z nich nie zdiagnozowano tego schorzenia.

Tabela 6.17: Tabela liczności dla zmiennej *vf* ze zbioru danych `tTRACE`

Poziom	Liczba obserwacji
0	929
1	71

Podobnie jak w przypadku cukrzycy, migotanie komór dotyczy mniejszości pacjentów, bowiem zgodnie z wynikami przedstawionymi w Tabeli 6.17 u 929 z nich nie stwierdzono tej choroby. Większość chorych jest zatem w dość dobrym stanie.

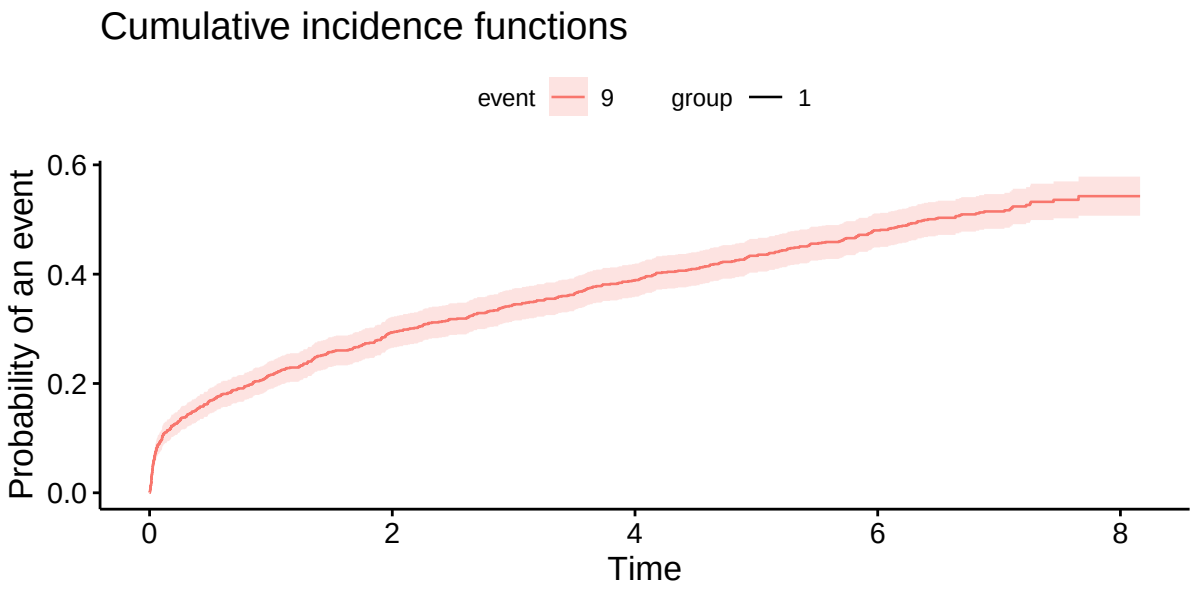
Na Rysunku 6.16 przedstawiliśmy histogram liczności obserwacji z uwzględnieniem cenzurowania.



Rysunek 6.16: Histogram czasu przeżycia dla zbioru danych `tTRACE`

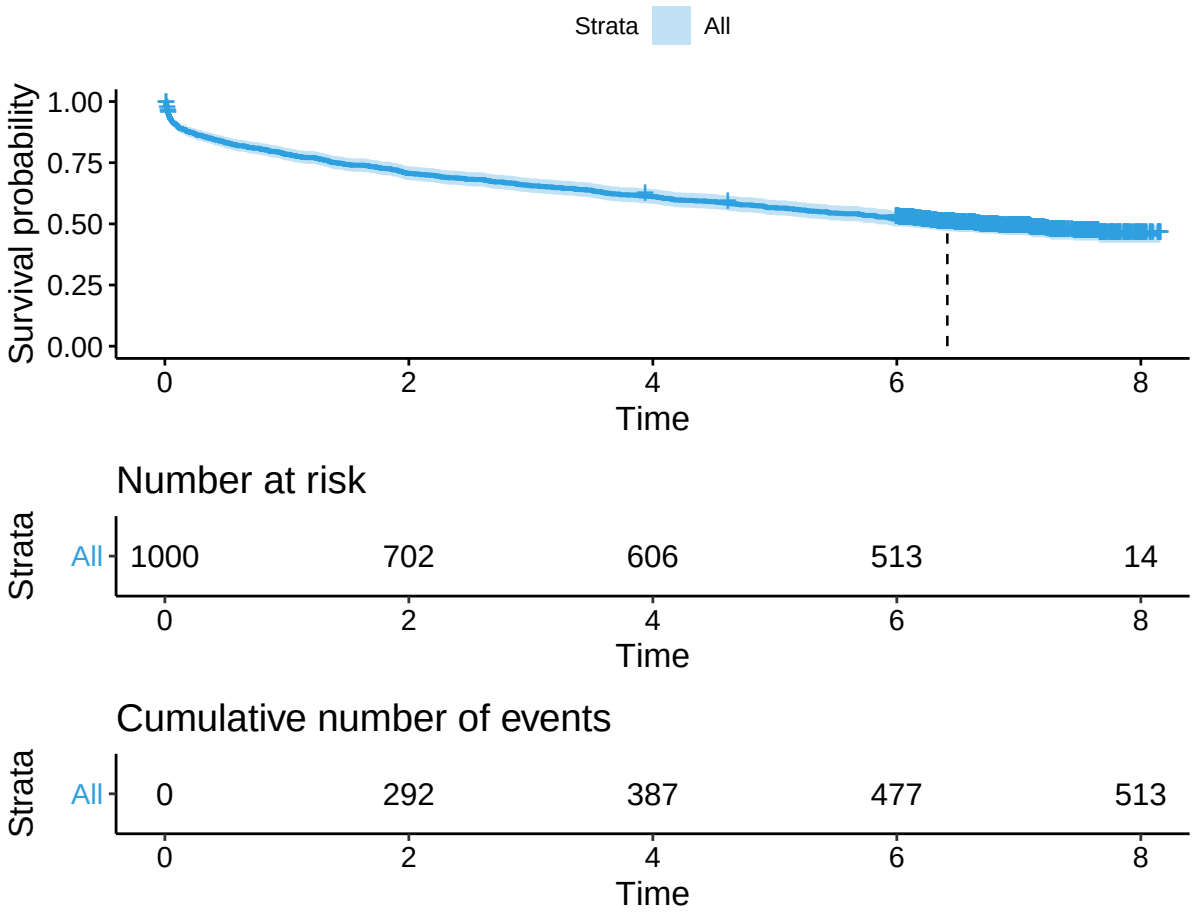
Łatwo widać, że czas przeżycia dla obserwacji cenzurowanych jest zdecydowanie dłuższy. Wśród wszystkich 1000 obserwacji liczba obserwacji cenzurowanych wynosi 487, co stanowi frakcję 48.7%.

Na Rysunku 6.17 przedstawiliśmy skumulowaną częstość wystąpienia zdarzenia.



Rysunek 6.17: Skumulowana częstość występowania zdarzenia dla zbioru danych `tTRACE`

Skumulowana częstość wystąpienia zdarzenia, jakim jest śmierć z powodu zawału mięśnia sercowego początkowo szybko wzrasta, lecz po około miesiącu wzrost ten staje się zdecydowanie wolniejszy. Prawdopodobieństwo wystąpienia zdarzenia przed upływem 8 lat wynosi około 0.6.



Rysunek 6.18: Estymowana funkcja przeżycia wraz z medianą czasu przeżycia dla zbioru danych `tTRACE`

Analiza estymowanej funkcji przeżycia zaprezentowanej na Rysunku 6.18 prowadzi do wniosku, że początkowo ryzyko śmierci szybko wzrasta, lecz po niedługim okresie czasu prawdopodobieństwo przeżycia maleje zdecydowanie wolniej. Obserwacje są zatem analogiczne do tych, które pojawiły się przy interpretacji wykresu skumulowanej częstości występowania.

W kolejnym kroku przedstawimy procedurę dopasowania modelu ze współczynnikiem zależnym od czasu. W kroku początkowym jednak zbudujemy model proporcjonalnych hazardów Coxa przy wykorzystaniu wszystkich predyktorów. Pozwoli to stwierdzić, że w przypadku danych tu rozpatrywanych model ten nie jest odpowiedni.

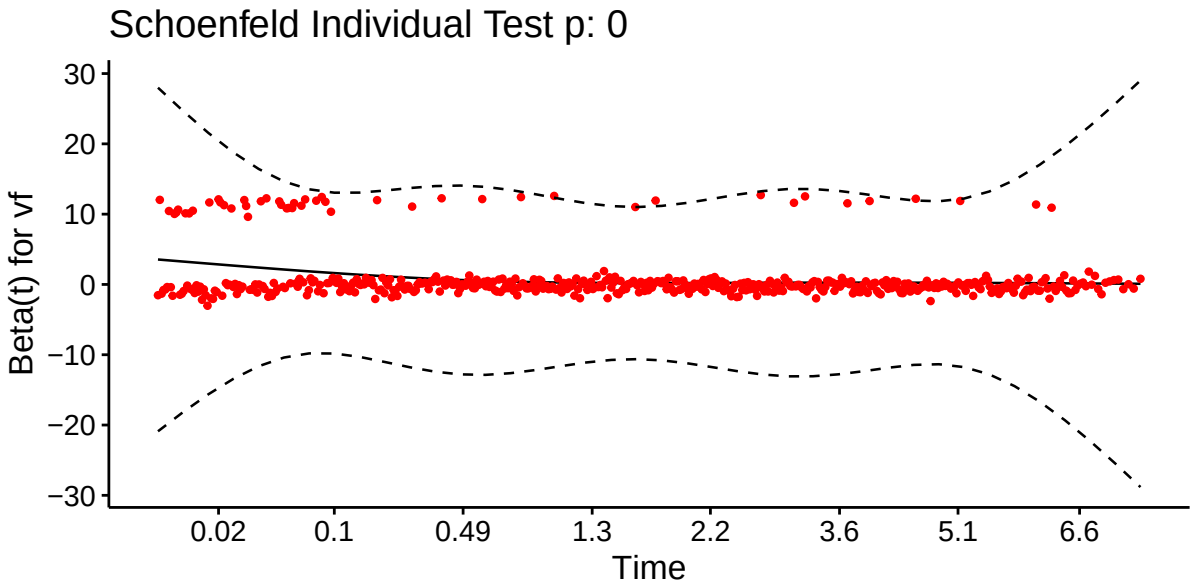
```
cox.full<-coxph(Surv(time,status==9)~chf+diabetes+vf+wmi+sex+age,
                data=tTRACE)
```

Po utworzeniu modelu przejdziemy do weryfikacji założenia o proporcjalności hazardów. Wyniki otrzymane po zastosowaniu funkcji `cox.zph()` są widoczne w Tabeli 6.18.

Tabela 6.18: Wyniki uzyskane w teście weryfikującym proporcjonalność hazardów dla danych `tTRACE`

	Wartość statystyki	Liczba stopni swobody	Wartość p
chf	0.83	1.00	0.36
diabetes	0.72	1.00	0.40
vf	23.01	1.00	0.00
wmi	2.20	1.00	0.14
sex	0.12	1.00	0.73
age	0.30	1.00	0.58
GLOBAL	25.72	6.00	0.00

Wartości p zawarte w Tabeli 6.18 pozwalają stwierdzić, że model proporcjonalnych hazardów Coxa nie jest dobrze dopasowany. Należy bowiem odrzucić hipotezę, że hazard jest proporcjonalny w grupach pacjentów chorujących na migotanie komór oraz niecierpiących na to schorzenie. Wynik otrzymany dla tej zmiennej zilustrowano na Rysunku 6.19.



Rysunek 6.19: Rezydua Schoenfelda dla zmiennej *vf*

Rysunek 6.19 również świadczy o tym, że założenie proporcjonalności hazardów względem zmiennej *vf* nie jest spełnione, w szczególności w pierwszym miesiącu prowadzenia badań. W takiej sytuacji zastosowanie znajdzie zatem model ze współczynnikiem zależnym od czasu. Dopasowanie takiego modelu umożliwia funkcja `timecox()` z pakietu `timereg`.

```
tcc.full<-timecox(Surv(time,status==9)~chf+diabetes+vf+wmi+sex+age,
              data=tTRACE,n.sim=100)
```

Model ten jest dopasowywany przy użyciu metod resamplingu. Z tego powodu zachodzi konieczność ustawienia liczby symulacji. W naszych analizach parametr `n.sim` odpowiadający za ich liczbę wynosi sto. Pozwoli nam to otrzymać dość dobrą dokładność i jednocześnie uniknąć zbyt długiego czasu obliczeń.

Funkcja `summary()` zwraca wyniki trzech testów. Pierwszy z nich jest testem istotności opartym na statystyce supremum (ang. *supremum-test of significance*). Wartości statystyki testowej i poziomów krytycznych dla poszczególnych zmiennych uzyskanych przy użyciu tego testu zawarte są w Tabeli 6.19. Weryfikuje on hipotezę zakładającą, że poszczególne współczynniki modelu są równe zero. Kolejne dwa z nich to testy Kołmogorowa-Smirnowa oraz Cramera von Misesa. W obu tych testach hipoteza zerowa stanowi, że współczynniki modelu nie zależą od czasu. Wyniki tychże testów zostały przedstawione odpowiednio w Tabelach 6.20 oraz 6.21.

Tabela 6.19: Wyniki uzyskane w teście istotności opartym na statystyce supremum dla pełnego modelu ze współczynnikiem zależnym od czasu

	Wartość statystyki testowej	Wartość p
(Intercept)	11.30	0.00
chf1	5.38	0.00
diabetes1	3.03	0.03
vf1	4.28	0.00
wmi	6.25	0.00
sex1	1.76	0.71
age	9.84	0.00

Korzystając z testu supremum, możemy stwierdzić, że jedyną zmienną nieistotną w modelu jest zmienna *sex*, ponieważ jest to jedyna zmienna, dla której nie ma podstaw do odrzucenia hipotezy zerowej. Reszta współczynników modelu jest bowiem istotna statystycznie.

Tabela 6.20: Wyniki uzyskane w teście Kołmogorowa-Smirnowa dla pełnego modelu ze współczynnikiem zależnym od czasu

	Wartość statystyki testowej	Wartość p
(Intercept)	6.11	0.10
chf1	0.61	0.88
diabetes1	0.60	0.94
vf1	1.94	0.12
wmi	1.32	0.39
sex1	0.79	0.80
age	0.07	0.24

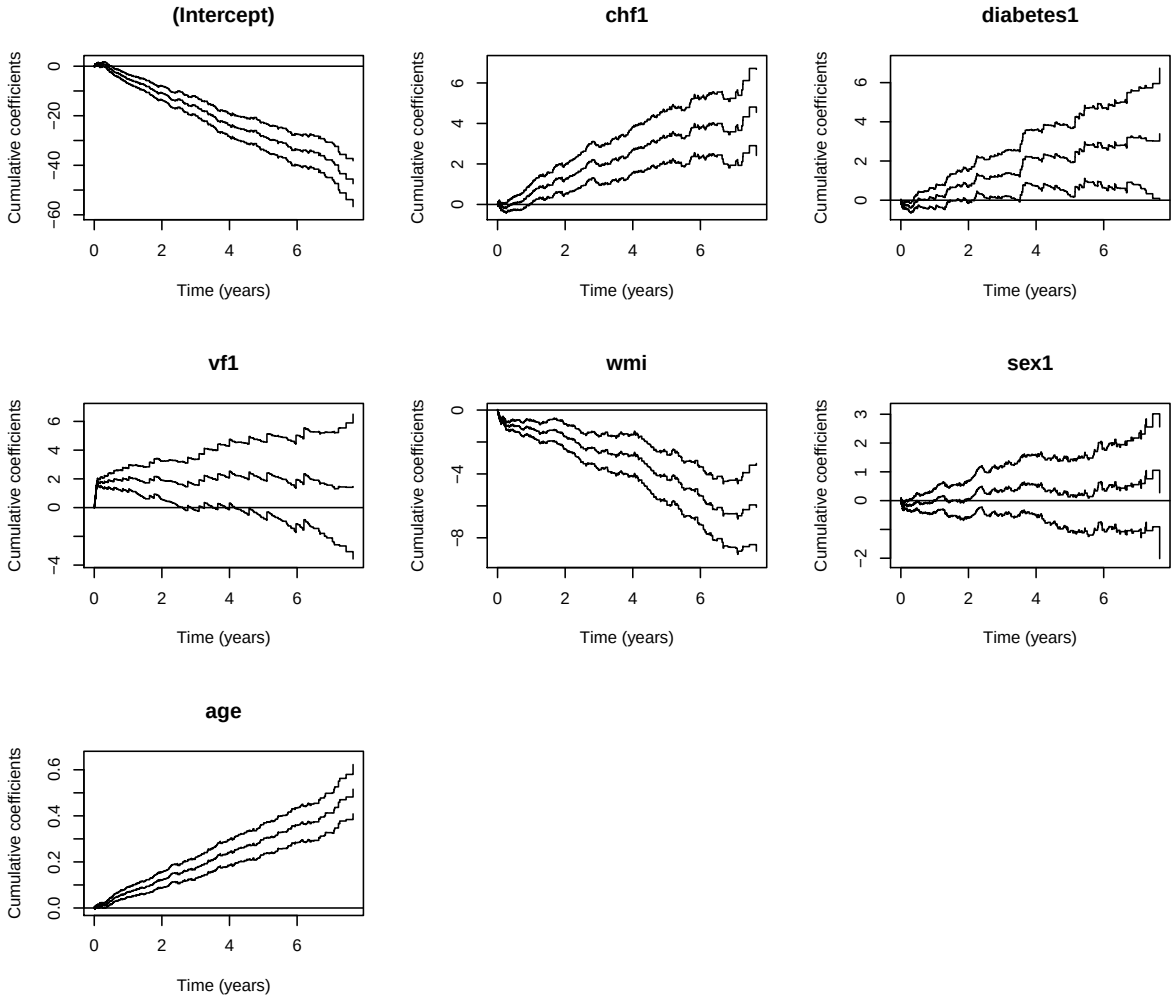
Korzystając z testu Kołmogorowa-Smirnowa, na poziomie istotności 0.05, nie ma podstaw do odrzucenia hipotezy, że współczynniki przy każdej ze zmiennych w rozpatrywanym modelu nie zależą od czasu. Warto dodać, że nie biorąc pod uwagę wyrazu wolnego, najmniejsza wartość p została osiągnięta dla zmiennej *vf*. Wiążące są jednak również wyniki testu Cramera von Misesa, ponieważ często wyniki obu tych testów mimo weryfikowania tej samej hipotezy, nie prowadzą do tych samych wniosków.

Tabela 6.21: Wyniki uzyskane w teście Cramera von Misesa dla pełnego modelu ze współczynnikiem zależnym od czasu

	Wartość statystyki testowej	Wartość p
(Intercept)	50.70	0.20
chf1	0.56	0.80
diabetes1	0.46	0.91
vf1	12.40	0.03
wmi	1.80	0.53
sex1	0.56	0.88
age	0.01	0.23

Korzystając z testu Cramera von Misesa na poziomie istotności 0.05, należy jednak odrzucić hipotezę, że współczynnik występujący przy zmiennej *vf1* nie zależy od czasu. W przypadku pozostałych zmiennych można przyjąć, że związane z nimi współczynniki nie zależą od czasu.

Kolejnym krokiem analizy jest zbadanie skumulowanych oszacowanych współczynników modelu określonych wzorem $\int_0^t \hat{\beta}(u)du$.



Rysunek 6.20: Estymowane skumulowane współczynniki wraz z 95% przedziałem ufności dla pełnego modelu ze współczynnikiem zależnym od czasu

Wykresy przedstawione na Rysunku 6.20 podsumowują wcześniej przeprowadzone rozważania. Prezentują one zależność wartości skumulowanych współczynników modelu od czasu wraz z 95% przedziałami ufności. Można zauważyć, że wartości skumulowanego współczynnika dla zmiennej *sex* przecinają krzywą poziomą, która w teście istotności opartym na statystyce supremum odpowiada hipotezie zerowej. Z tego powodu wcześniej nie było podstaw do jej odrzucenia. Co więcej, dla zmiennej *vf* widać, że wartości skumulowanego współczynnika wraz z przedziałami ufności nie są monotoniczne. Oznacza to, że współczynnik dla zmiennej związanej z migotaniem komór powinien zależeć od czasu. Warto przypomnieć, że do takiego wniosku prowadził też wynik testu Cramera von Misesa.

Po wykonaniu powyższych testów możemy przejść do budowy modelu ponownie przy wykorzystaniu funkcji `timecox()`, lecz z zastosowaniem dla wybranych predyktorów funkcji `const()`, która umożliwia sprecyzowanie, która ze zmiennych ma stały wraz z upływem czasu wpływ na model. Zgodnie z wynikami wcześniejszych analiz, zastosowaliśmy ją do wszystkich zmiennych z wyłączeniem zmiennej *vf*.

```
tcc.full.const<-timecox(Surv(time,status==9)~const(chf)+const(diabetes)+vf+
const(wmi)+const(sex)+const(age),data=tTRACE,n.sim=100)
```

Wyniki uzyskane w teście supremum są widoczne w Tabeli 6.22.

Tabela 6.22: Wyniki uzyskane w teście istotności opartym na statystyce supremum dla modelu `tcc.full.const`

	Wartość statystyki testowej	Wartość p
(Intercept)	68.20	0.00
vf1	4.17	0.00

Z Tabeli 6.22 wynika, że zarówno w przypadku zmiennej *vf* jak i wyrazu wolnego należy odrzucić hipotezę o nieistotności współczynników.

Następnie przejdziemy do analizy modelu obejmującej wartości jego współczynników wraz z wybranymi wartościami.

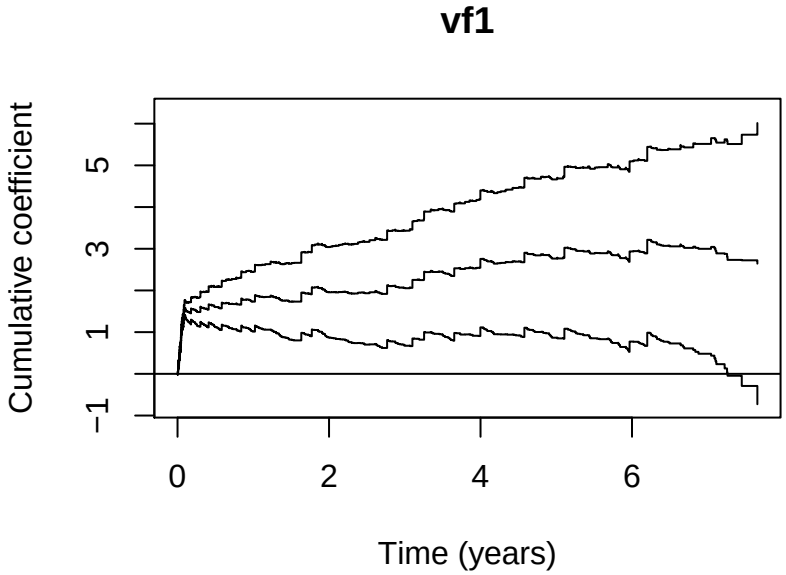
Tabela 6.23: Współczynniki wraz z wybranymi wartościami dla modelu `tcc.full.const`

	β	SE_β	odporny SE_β	z	wartość p	lower	upper
const(chf)1	0.59	0.10	0.11	5.51	0.00	0.38	0.80
const(diabetes)1	0.41	0.13	0.18	2.29	0.02	0.15	0.66
const(wmi)	-0.93	0.12	0.14	-6.47	0.00	-1.16	-0.70
const(sex)1	0.07	0.10	0.11	0.68	0.05	-0.12	0.26
const(age)	0.06	0.00	0.01	10.50	0.00	0.05	0.07

Z Tabeli 6.23 wynika, że dla wszystkich zmiennych znajdujących się w modelu `tcc.full.const` należy odrzucić hipotezę zerową o ich nieistotności. Wartości błędów standardowych dla poszczególnych współczynników SE_β są dość małe. Ponadto zauważmy, że wartości wszystkich współczynników poza związanym ze zmienną *wmi* są dodatnie co będzie miało wpływ na ich interpretację. Mianowicie wzrost miary efektywności pompowania krwi przez serce ma odmienny wpływ na ryzyko wystąpienia zdarzenia niż pozostałe predyktory.

Z powodu istotności wszystkich zmiennych znajdujących się w modelu pełnym oraz faktu, że nie ma możliwości wykorzystania wartości logarytmu funkcji wiarogodności, nie zostanie zastosowana żadna metoda krokowa wyboru zmiennych do modelu. Wybrany przez nas model jest zatem modelem ze wszystkimi predyktorami.

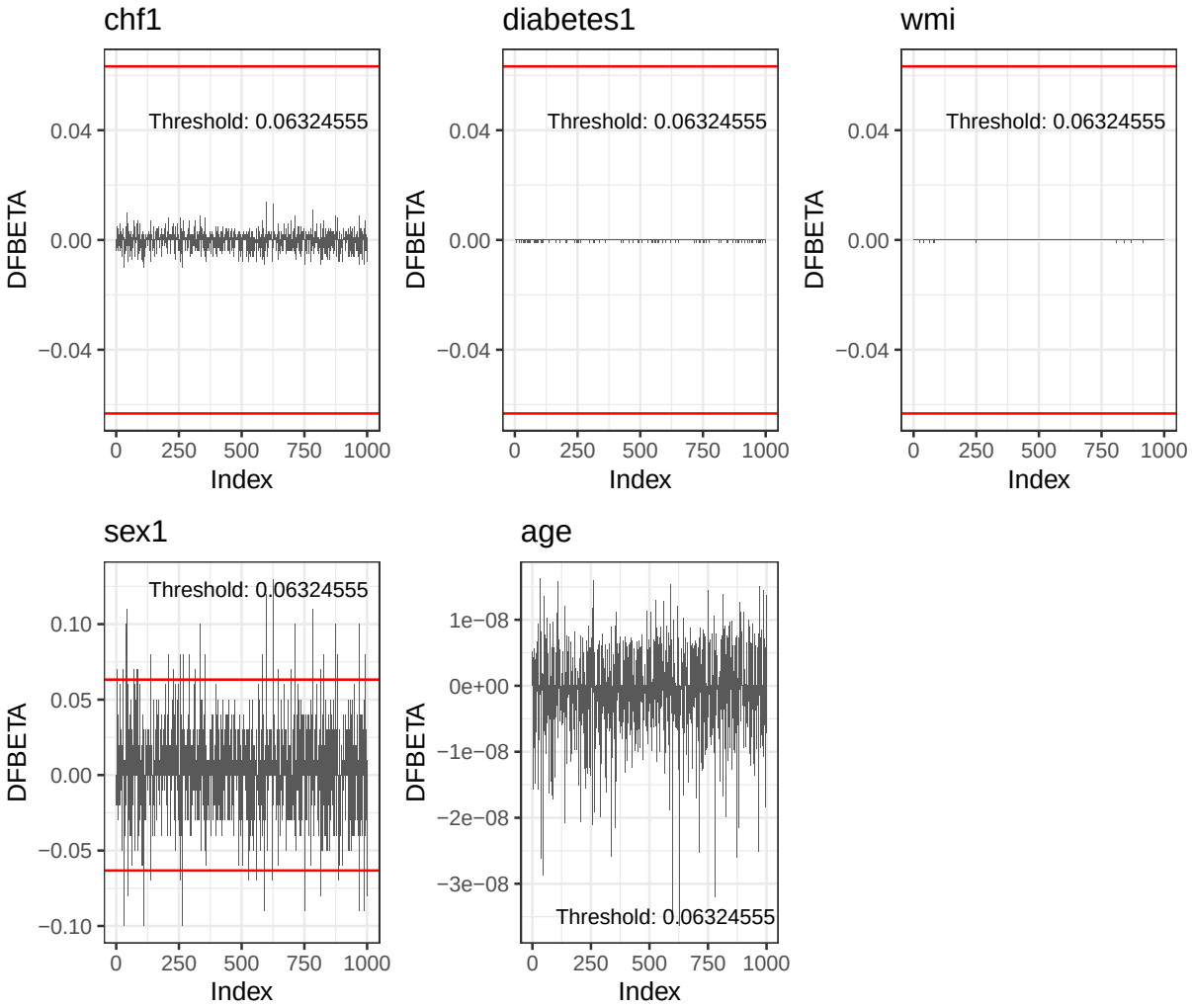
Przechodząc do dokładnej interpretacji współczynników modelu możemy zauważyć, że wszystkie predyktory za wyjątkiem *wmi* zwiększają ryzyko śmierci z powodu zawału serca. Kliniczna niewydolność pompy serca zwiększa ryzyko zgonu o ponad 80%, cukrzyca o około 50%, płeć żeńska o 7%, a upływ jednego roku o 6%. Zwiększenie efektywności pompowania krwi o jednostkę powoduje spadek ryzyka śmierci o ponad 60%. We wnioskowaniach tych zakłada się, że wartości pozostałych predyktorów są ustalone. Można podejrzewać, że największym czynnikiem ryzyka śmierci z powodu zawału mięśnia sercowego jest kliniczna niewydolność pompy serca. Należy jednak pamiętać, że w interpretacji pominęliśmy współczynnik związany z migotaniem komór, który zmienia się w czasie. Jego skumulowane wartości wraz z 95% przedziałem ufności przy założeniach modelu `tcc.full.const` przedstawiliśmy na Rysunku 6.21.



Rysunek 6.21: Estymowany skumulowany współczynnik wraz z 95% przedziałem ufności dla zmiennej *vf* i modelu `tcc.full.const`

Wyniki zaprezentowane na Rysunku 6.21 sugerują, że skumulowane oszacowane wartości współczynnika przy zmiennej *vf* skupiają się wokół wartości 2. Przedstawiony wykres nie jest niestety wykresem estymowanych wartości współczynnika, jednak niesie pewną informację o związku między czasem a daną charakterystyką.

W kolejnym kroku za pomocą wartości $DFBETA$ sprawdzimy, czy dla zbudowanego modelu możemy zidentyfikować obserwacje, które mają zauważalny wpływ na model.



Rysunek 6.22: Wartości $DFBETA$ dla modelu `tcc.full.const`

Wartości $DFBETA$ zostały przedstawione na Rysunku 6.22. Żadna z obserwacji nie ma dużego wpływu na wartość współczynnika zmiennych *chf*, *diabetes*, *wmi* oraz *age*, ponieważ otrzymane wartości $DFBETA$ mieszczą się w przedziale $[-0.06324555, 0.06324555]$. Jednakże dla zmiennej *sex* w przypadku obserwacji o numerach 2, 26, 32, 39, 44, 48, 70, 84, 86, 109, 137, 139, 208, 227, 256, 257, 262, 267, 291 , 316, 335, 355, 529, 561, 584, 589, 597, 621, 628, 647, 698, 712, 751, 782, 814, 815, 826, 874, 883, 887, 967, 970, 989, 994 oraz 1000 obserwujemy wartości $DFBETA$ przekraczające próg. Ich liczba wynosi 45, lecz należy pamiętać, że liczba obserwacji w zbiorze jest dość duża i wynosi 1000. Warto zaznaczyć, że istnieje wiele strategii postępowania z tego typu obserwacjami. Zawsze należy je usuwać, gdy ma się pewność, że zostały one wprowadzone do zbioru danych w sposób nieprawidłowy. Tutaj jednak nie ma możliwości zweryfikowania tego faktu ze względu na różnorodność i mnogość zmiennych objaśniających. Kolejną przesłanką do usunięcia obserwacji zakwalifikowanych jako wpływowe może być naruszenie założeń modelu. Jednak jak wynika z wcześniej przeprowadzonych analiz, ostateczny model cechuje się dobrymi własnościami. Co więcej, przekroczenie progu ma miejsce w przypadku tylko jednego z pięciu współczynników. Zadecydowaliśmy zatem aby nie usuwać ze zbioru zidentyfikowanych obserwacji.

6.3 Model proporcjonalnych szans

Kolejnym modelem, dla którego przeprowadziliśmy analizy jest model proporcjonalnych szans. Wybrane dane dotyczą pacjentów chorych na przewlekłą białaczkę szpikową. Można je znaleźć pod nazwą `wbc1` w pakiecie `dynpred`. Badanie zostało przeprowadzone w Beneluksie w 1998 roku. Wymiar ramki danych to 190×5, obejmuje ona zatem 190 pacjentów oraz 5 zmiennych. Poniżej zamieściliśmy szczegółowy opis tychże zmiennych.

- *patnr* (typ: *numeric*) – ID pacjenta,
- *tyears* (typ: *numeric*) – czas w latach do śmierci lub ostatniej kontroli lekarskiej,
- *d* (typ: *numeric*) – status pacjenta (śmierć=1, cenzura=0),
- *sokal* (typ: *numeric*) – wskaźnik Sokala zależący od rozmiaru śledziony, udziału procentowego komórek blastycznych, płytek krwi oraz wieku w momencie diagnozy,
- *age* (typ: *numeric*) – wiek w momencie diagnozy.

W zbiorze danych `wbc1` nie zidentyfikowaliśmy żadnych brakujących obserwacji. Co więcej, zgodnie z powyższym opisem, wszystkie zmienne objaśniające są odpowiedniego typu.

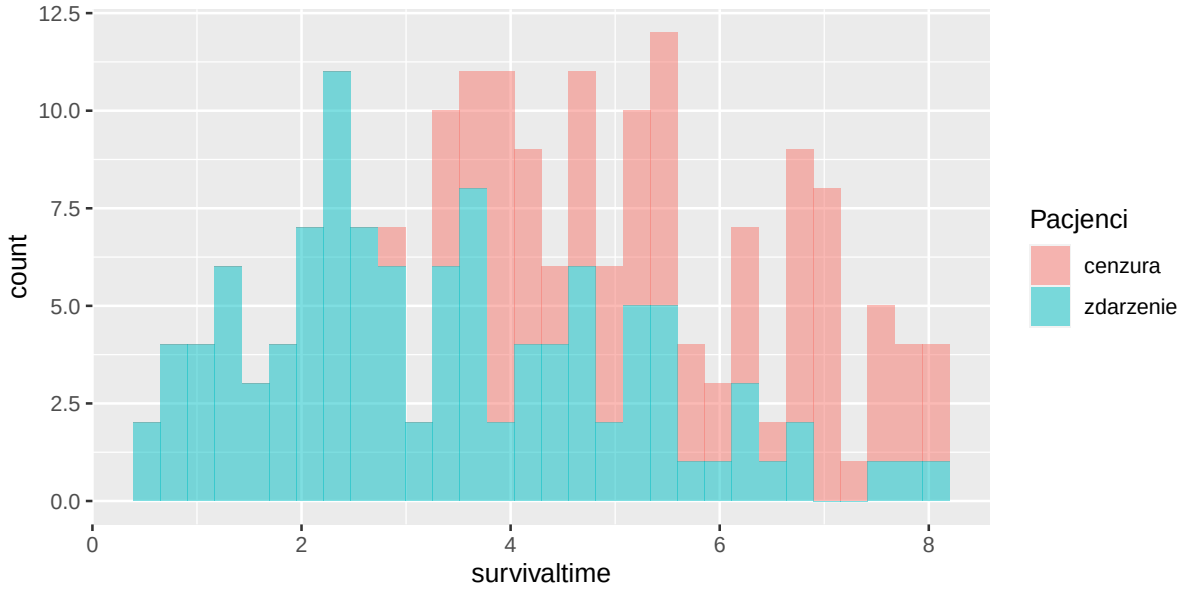
Następnie obliczyliśmy wartości podstawowych wskaźników statystycznych dla wszystkich zmiennych ciągłych. Wyniki zostały zawarte w Tabeli 6.24.

Tabela 6.24: Wartości wskaźników dla zmiennych ciągłych ze zbioru danych `wbc1`

	sokal	*age*
Średnia	1.07	53.41
Wariancja	0.26	172.69
Odchylenie standardowe	0.51	13.14
Mediana	0.95	55.75
Pierwszy kwartyl	0.75	43.35
Trzeci kwartyl	1.30	64.18
Minimum	0.32	19.90
Maksimum	4.44	84.20

Wyniki zaprezentowane w Tabeli 6.24 pozwalają zauważyć, że średni wiek pacjentów wynosił ponad 50 lat. Najmniejsze zróżnicowanie danych ma miejsce w przypadku predyktora *sokal*. 25% pacjentów miało wskaźnik Sokala mniejszy bądź równy 0.75 oraz nie osiągnęła bądź skończyła 43.35 rok życia. Oczywiście analogiczna iterpretacja ma miejsce w przypadku mediany i trzeciego kwartyla, których wartości dotyczą odpowiednio frakcji populacji równej 50% oraz 75%. Wartości rozważanych zmiennych należą odpowiednio do przedziałów $[0.32, 4.44]$ oraz $[19.9, 84.2]$.

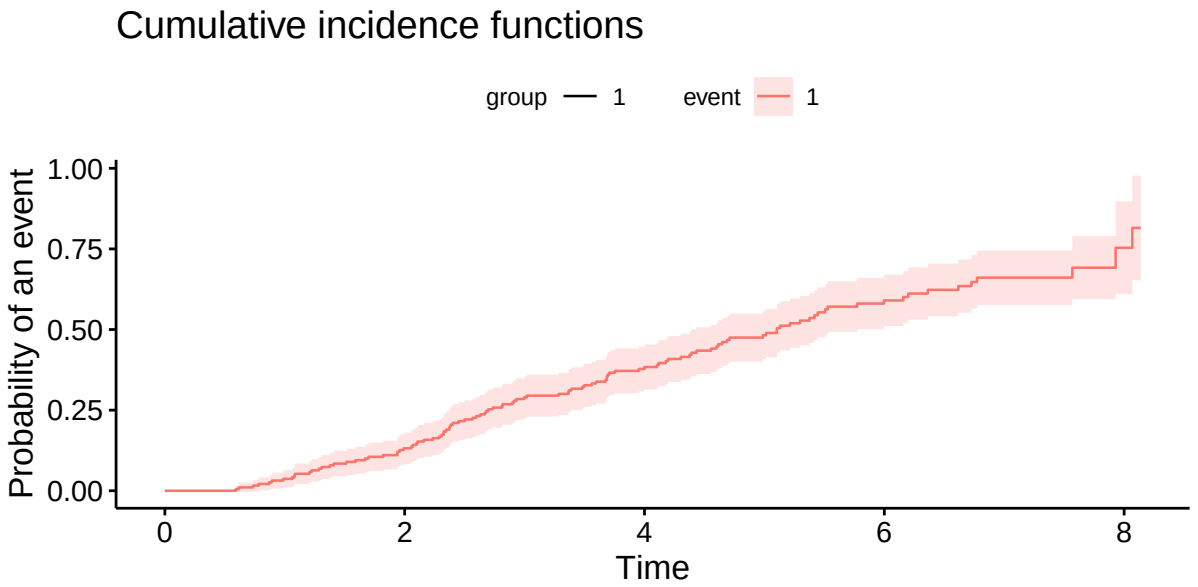
Na Rysunku 6.23 przedstawiliśmy histogram liczności obserwacji dla zmiennej *d*.



Rysunek 6.23: Histogram czasu przeżycia dla zbioru danych `wbc1`

Z Rysunku 6.23 możemy wnioskować, że liczność obserwacji cenzurowanych oraz tych, dla których zaszło zdarzenie jest podobna. Przechodząc do dokładnych danych, liczba pacjentów, dla których nie dysponujemy kompletną informacją wynosi 81, natomiast dla 109 pacjentów zaobserwowano zdarzenie. Frakcja obserwacji cenzurowanych jest zatem dość duża i wynosi 42.63%.

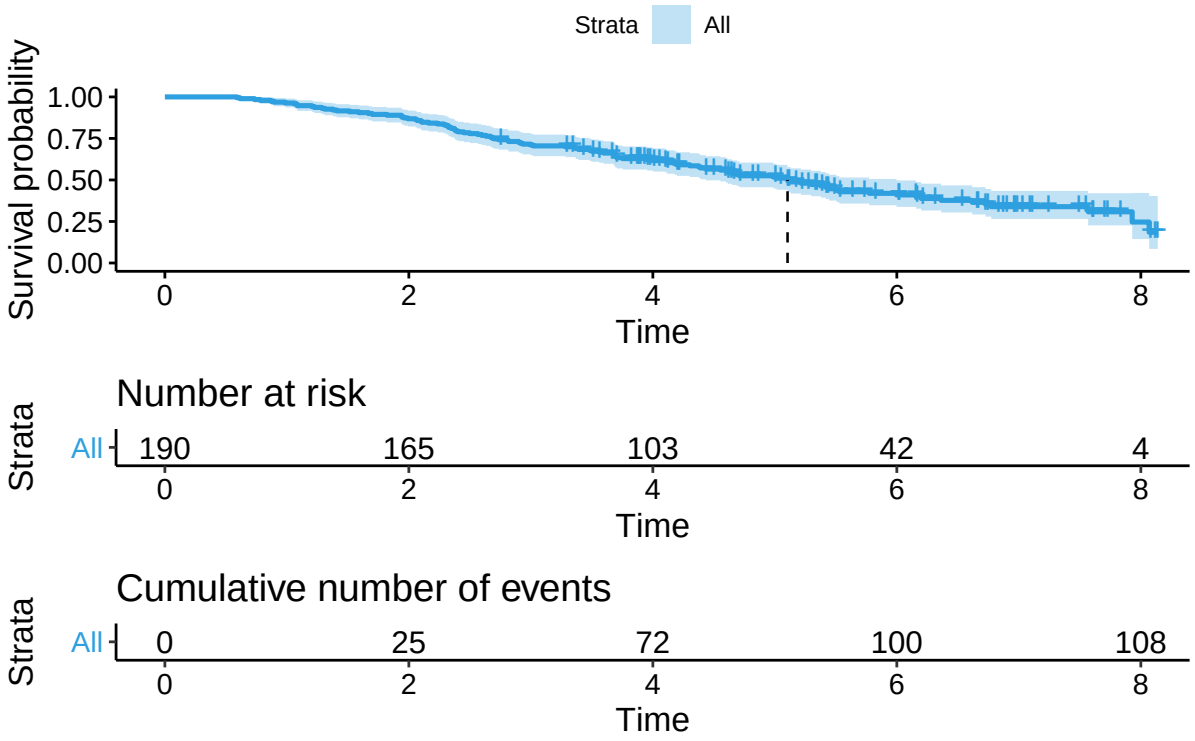
Przeanalizowaliśmy również skumulowaną częstość występowania zdarzenia. Wyniki przedstawiliśmy na Rysunku 6.24.



Rysunek 6.24: Skumulowana częstość występowania zdarzenia dla zbioru danych `wbc1`

Na podstawie uzyskanych wyników możemy wywnioskować, że prawdopodobieństwo wystąpienia zdarzenia przed upływem 8 lat wynosi około 0.75. Oznacza to, że po upływie tego okresu czasu pacjent umrze lub odbędzie wizytę lekarską z takim właśnie praw-

dopodobieństwem. Warto dodać, że prawdopodobieństwo to wzrasta w stałym tempie.



Rysunek 6.25: Estymowana funkcja przeżycia wraz z medianą czasu przeżycia dla zbioru danych `wbc1`

Na Rysunku 6.25 przedstawiliśmy estymowaną funkcję przeżycia wraz z przedziałem ufności. Widać, że spadek prawdopodobieństwa przeżycia nie jest bardzo szybki. Oznacza to, że pacjenci najczęściej umierają lub odbywają wizytę lekarską po dłuższym okresie czasu.

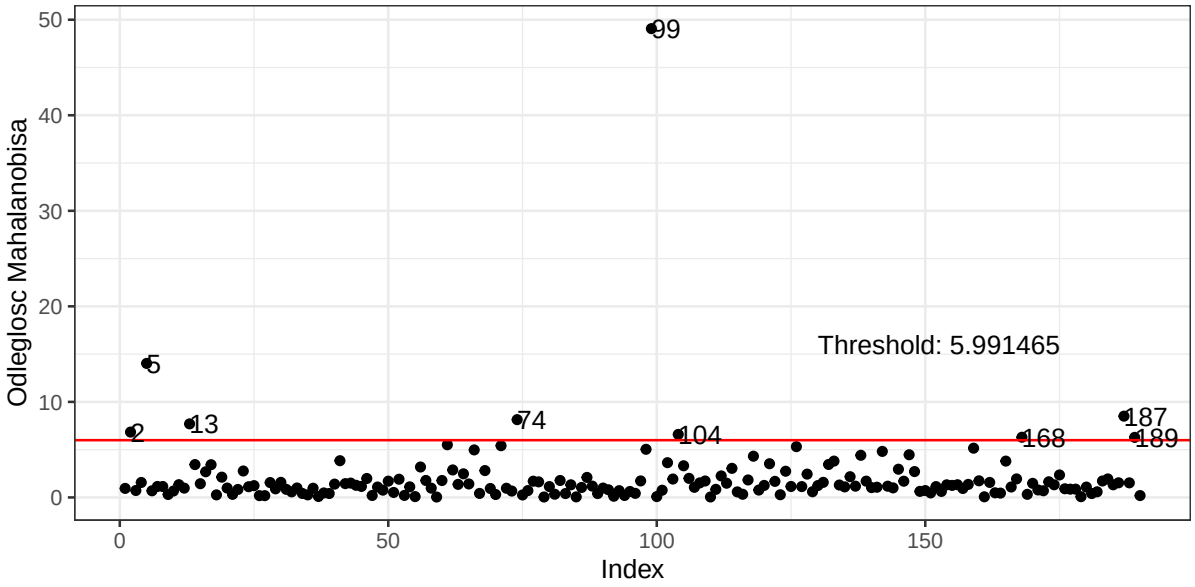
Z uwagi na fakt, że wszystkie zmienne objaśniane w zbiorze danych `wbc1` są ilościowe, w celu dokładniejszego zapoznania się z danymi, dokonamy identyfikacji obserwacji odstających ze względu na odległość Mahalanobisa.

```
matrix.wbc1<-wbc1[,4:5]
m.distances.wbc1<-mahalanobis(matrix.wbc1,colMeans(matrix.wbc1),cov(matrix.wbc1))
```

Zgodnie z wcześniejszym opisem teoretycznym, zastosowany próg będzie równy $\chi^2_2(1-\alpha)$, gdzie $\alpha = 0.05$.

```
p=2
threshold.wbc1<-qchisq(0.95, df=p, ncp = 0, lower.tail = TRUE, log.p = FALSE)
```

Dla danych `wbc1` próg ten wynosi 5.991465. Wyniki zostały zilustrowane na Rysunku 6.26.



Rysunek 6.26: Odległość Mahalanobisa dla zbioru danych `wbc1`

Wcześniej wspomniana wartość progu została zaznaczona na Rysunku 6.26 poziomą czerwoną linią. Wynika z niego, że obserwacje o numerach 2, 5, 13, 74, 99, 104, 168, 187 oraz 189 powinny zostać uznane za odstające ze względu na wartości zmiennych objaśniających. Nie jesteśmy jednak w stanie stwierdzić, czy zaszła pomyłka podczas wprowadzania danych. Przyjrzyjmy się zakresom wartości osiąganych przez zmienne *sokal* oraz *age*, które zostały już wcześniej przedstawione w Tabeli 6.24. Oczywiście wiek pacjentów należący do przedziału $[19.9, 84.2]$ jest jak najbardziej prawdopodobny. Z kolei prawidłowość zakresu wyników osiąganych dla wskaźnika Sokala nie jest łatwa do zweryfikowania natychmiast; powinna być ona skonsultowana ze specjalistą. Na początku decydujemy pozostawić zidentyfikowane obserwacje w zbiorze. Zostaną one przez nas usunięte, jeśli po zbudowaniu modelu proporcjonalnych szans przy użyciu całego zbioru danych okaże się, że założenia modelu nie są spełnione.

Po wstępnym zapoznaniu się z danymi przy pomocy funkcji `prop.odds()` zbudujemy model proporcjonalnych szans przy wykorzystaniu wszystkich predyktorów.

```
fit.po.full<-prop.odds(Event(tyears,d==1)~sokal+age,data=wbc1)
```

Funkcja `summary()` zwraca informacje na temat wyników różnych testów dotyczących modelu proporcjonalnych szans. Pierwszym z nich jest test istotności dla wyrażenia $G(t)$ oparty na statystyce supremum (ang. *supremum-test of significance*). Otrzymując wynik 0.344 należy stwierdzić, że nie ma podstaw do odrzucenia hipotezy o nieistotności tego wyrażenia. Kolejnym testem jest test oparty na statystyce Kołmogorowa-Smirnowa. Hipoteza zerowa w tym teście zakłada, że zmienne nie zależą od czasu. Wartość p równa 0.332 pozwala stwierdzić, że nie ma podstaw do odrzucenia tejże hipotezy.

W Tabeli 6.25 przedstawiliśmy pozostałe wartości zwracane przez funkcję `summary()`.

Tabela 6.25: Współczynniki wraz z wybranymi wartościami dla pełnego modelu proporcjonalnych szans

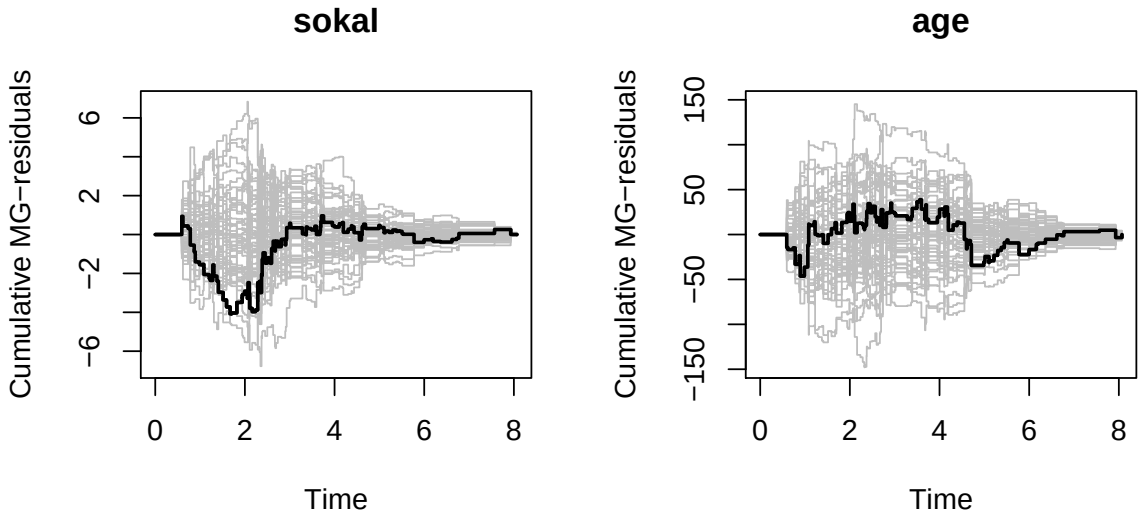
	współczynnik	SE	odporny SE	$D2\log(L)^{-1}$	z	wartość p	lower	upper
sokal	0.63	0.24	0.21	0.25	2.97	0.00	0.16	1.10
age	0.02	0.01	0.01	0.01	2.12	0.03	0.00	0.04

Łatwo widać, że dla wszystkich zmiennych znajdujących się w modelu należy odrzucić hipotezę o nieistotności. Co więcej, możemy przypuszczać, że zmienną, która jest najbardziej warta uwzględnienia w modelu jest zmienna *sokal*.

Tabela 6.26: Wyniki testu *Goodness-of-fit* dla pełnego modelu proporcjonalnych szans

	Statystyka testowa	wartość p
sokal	4.09	0.20
age	46.40	0.91

Ostatnim testem jest test dobroci dopasowania (ang. *Test of Goodness-of-fit*). Weryfikuje on, czy założenie proporcjonalnych szans jest spełnione. Wyniki przedstawione w Tabeli 6.26 nakazują stwierdzić, że nie ma podstaw do odrzucenia hipotezy o proporcjonalności szans. Zbudowany model jest zatem dobrze dopasowany. Z tego powodu zdecydydowaliśmy nie usuwać obserwacji, które wcześniej zidentyfikowaliśmy jako odstające przy pomocy odległości Mahalanobisa.



Rysunek 6.27: Reszty skumulowane dla pełnego modelu proporcjonalnych szans

Na Rysunku 6.27 przedstawiliśmy skumulowane reszty martyngałowe dla pełnego modelu proporcjonalnych szans. Otrzymane wyniki pokrywają się z tymi, które otrzymaliśmy przy pomocy testu dobroci dopasowania co jest równoznaczne z faktem, że założenie proporcjonalnych szans jest spełnione.

Mimo dobrego dopasowania modelu pełnego sprawdzimy, czy nie zostanie uzyskana poprawa uzyskanych wyników w przypadku eliminacji zmiennych. Podobnie jak w przypadku

modelu proporcjonalnych hazardów Coxa skorzystamy z metod krokowych opierających się na wartościach AIC oraz BIC. Wykonana zostanie jedynie eliminacja wsteczna. Wyniki zostały przedstawione w Tabeli 6.27.

Tabela 6.27: Wartości AIC i BIC dla modeli uzyskane przy pomocy eliminacji wstecznej dla modelu proporcjonalnych szans

	AIC	BIC
Model pełny	1004.00	1009.39
Model bez zmiennej *sokal*	1008.12	1010.81
Model bez zmiennej *age*	1006.40	1009.09

Analizując otrzymane wartości AIC oraz BIC należy stwierdzić, że najlepszym modelem jest model pełny w przypadku zastosowania kryterium AIC, oraz model bez zmiennej *age* w przypadku kryterium BIC. Otrzymane wyniki skłaniają zatem do przyjrzenia się własnościom modelu po usunięciu zmiennej *age*, który będzie jednocześnie modelem ostatecznie przyjętym.

```
fit.po.BIC<-prop.odds(Event(tyears,d==1)~sokal,data=wbc1)
```

Po utworzeniu odpowiedniego modelu ponownie przeanalizujemy wyniki otrzymane w testach, które są wykonywane po wywołaniu funkcji `summary()`.

Wartość p równa 0.006 uzyskana w teście istotności opartym na statystyce supremum nakazuje odrzucić hipotezę zerową o niestotności wyrażenia $G(t)$. Zauważmy, że otrzymaliśmy zatem poprawę, ponieważ w przypadku modelu pełnego nie było podstaw do odrzucenia tejże hipotezy. Przechodząc do analizy wyników otrzymanych w teście Kołmogorowa-Smirnowa, podobnie jak w przypadku modelu ze wszystkimi predyktorami, należy stwierdzić, że nie ma podstaw do odrzucenia hipotezy zerowej zakładającej, że zmienne nie zależą od czasu. Otrzymana wartość p w tym teście jest bowiem równa 0.264.

Tabela 6.28: Współczynniki wraz z wybranymi wartościami dla wybranego modelu proporcjonalnych szans

	β	SE_β	odporny SE	$D2\log(L)^{-1}$	z	wartość p	lower	upper
sokal	0.76	0.23	0.22	0.24	3.50	0.00	0.31	1.21

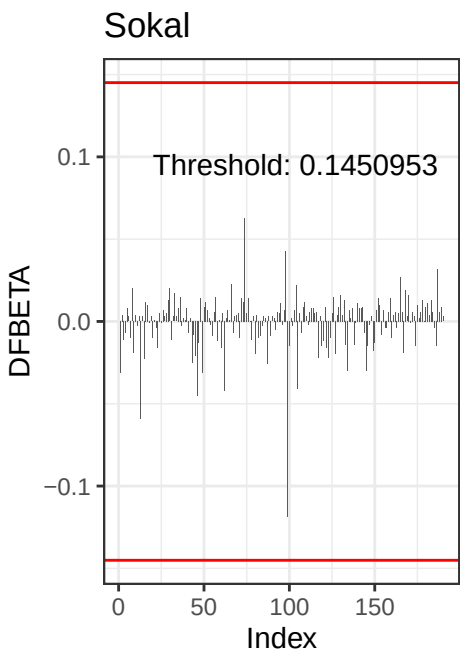
W modelu bez zmiennej *age*, dla predyktora *sokal* nadal powinniśmy odrzucić hipotezę zerową o nieistotności współczynnika. Wyniki, dzięki którym otrzymaliśmy taki wniosek zostały zawarte w Tabeli 6.28. Wartość współczynnika równa 0.76 skłania z kolei do wniosku, że w chwili początkowej ryzyko względne wynosi $e^{0.76}$. Oznacza to, że przy wzroście wartości wskaźnika Sokala o jednostkę, ryzyko wystąpienia zdarzenia wzrasta o ponad 120%.

Tabela 6.29: Wyniki testu *Goodness-of-fit* dla wybranego modelu proporcjonalnych szans

	Statystyka testowa	wartość p
sokal	4.45	0.19

Opierając się na wynikach zamieszczonych w Tabeli 6.29 otrzymujemy, że wybrany model jest dobrze dopasowany i spełnia założenie proporcjonalnych szans o czym świadczy wartość p otrzymana w teście *Goodness-of-fit*.

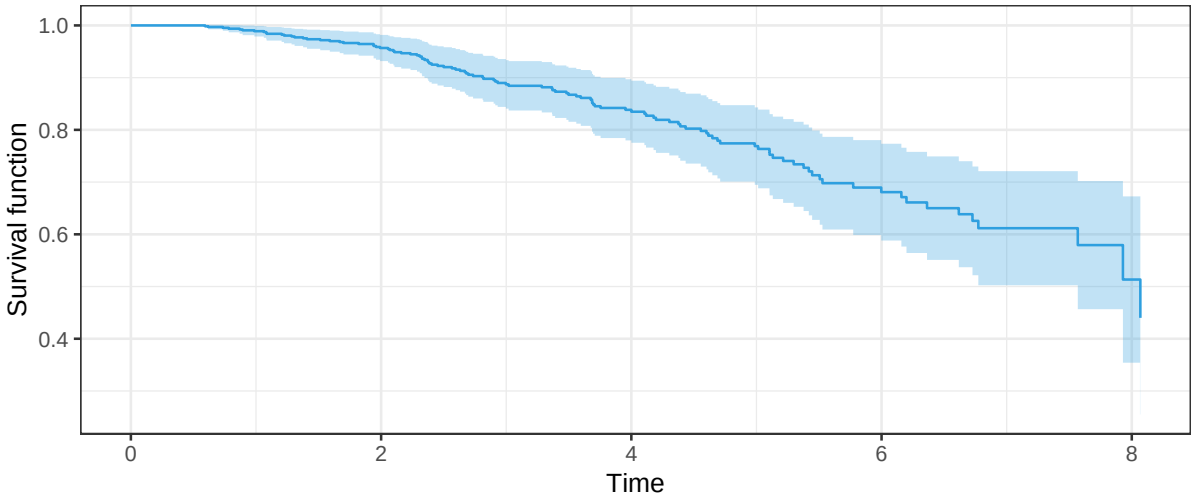
W celu diagnostyki utworzonego modelu przejdziemy do analizy wartości $DFBETA$. W tym przypadku są to różnice pomiędzy wartością współczynnika *sokal*, a tą samą wartością, lecz po usunięciu i-tej obserwacji. Dla danych `wbc1`, które zawierają informacje na temat 190 obserwacji, próg klasyfikacji obserwacji jako wpływowej wynosi 0.1450953.



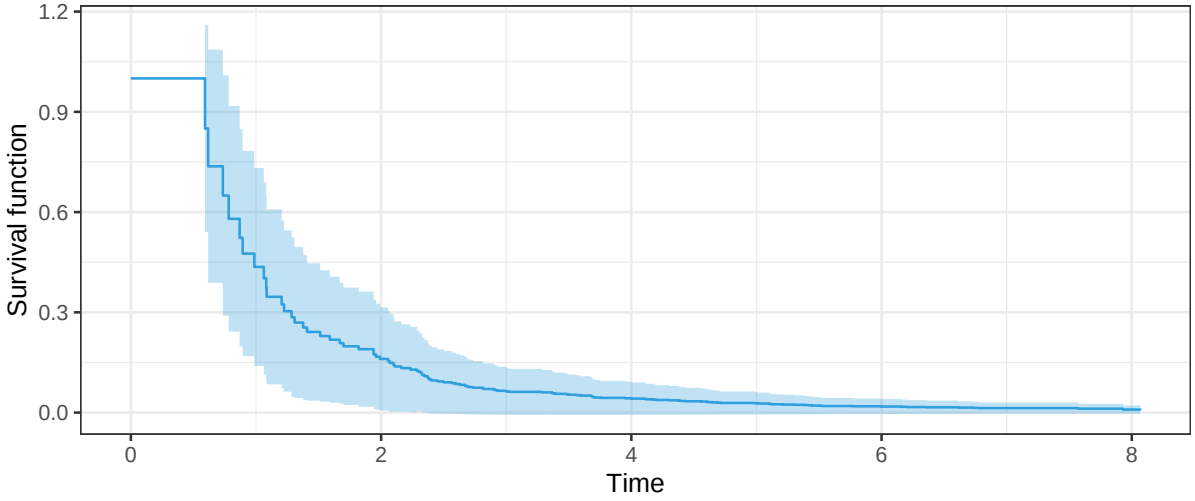
Rysunek 6.28: Wartości $DFBETA$ dla wybranego modelu proporcjonalnych szans

Rysunek 6.28 przedstawia otrzymane wartości $DFBETA$ dla wybranego modelu proporcjonalnych szans. Liniami czerwonymi zaznaczono wartości progowe. Łatwo widać, że nie mamy do czynienia z żadnymi obserwacjami wpływowymi, ponieważ wszystkie wartości $DFBETA$ mieszczą się w przedziale $[-0.1450953, 0.1450953]$. Wnioskujemy zatem, że żadna z obserwacji nie ma istotnego wpływu na wartości współczynników modelu.

Na Rysunkach 6.29 oraz 6.30 przedstawiono prognozowane przez model proporcjonalnych szans prawdopodobieństwa przeżycia dla pacjentów ze wskaźnikiem Sokala równym 0.5 oraz 4.



Rysunek 6.29: Prognozowane prawdopodobieństwo przeżycia dla pacjentów ze wskaźnikiem Sokala równym 0.5



Rysunek 6.30: Prognozowane prawdopodobieństwo przeżycia dla pacjentów ze wskaźnikiem Sokala równym 4

Rysunki 6.29 oraz 6.30 potwierdzają wcześniejsze obserwacje. Przy założeniach wybranego modelu proporcjonalnych szans zdecydowanie większe szanse na przeżycie mają pacjenci, u których wskaźnik Sokala jest niższy. Możemy podejrzewać, że wartość tego wskaźnika dobrze przewiduje prawdopodobieństwo przeżycia.

6.4 Model ryzyka addytywnego

Kolejnym modelem, dla którego przeprowadziliśmy analizę jest model ryzyka addytywnego. Zostanie on zbudowany dla danych `GBSG` z pakietu `mfp`. Wybrana ramka danych zawiera informacje na temat kobiet chorych na złośliwy nowotwór piersi, a dane zostały zebrane podczas badania *German Breast Cancer Study Group*. Dotyczą one 686 kobiet i zawierają 11 zmiennych. Poniżej zamieściliśmy ich krótki opis.

- *id* (typ: *integer*) – ID pacjentki,
- *htreat* (typ: *factor*) – indykator terapii hormonalnej,
- *age* (typ: *integer*) – wiek pacjentki w latach,
- *menostat* (typ: *factor*) – status menopauzalny (okres przedmenopauzalny=1, okres pomenopauzalny=2),
- *tumsize* (typ: *integer*) – rozmiar guza w *mm*,
- *tumgrad* (typ: *factor*) – stopień zaawansowania nowotworu, występuje na poziomach 1, 2 oraz 3,
- *posnodal* (typ: *integer*) – liczba węzłów chłonnych zajętych przez nowotwór,
- *prm* (typ: *integer*) – liczność receptorów progesteronowych wyrażona w femtomolach,
- *esm* (typ: *integer*) – liczność receptorów estrogenowych wyrażona w femtomolach,
- *rfst* (typ: *integer*) – czas wolny od nawrotów nowotworu wyrażony w dniach,
- *cens* (typ: *integer*) – indykator cenzury (zdarzenie=1, cenzura=0).

Zbiór danych został sprawdzony pod kątem obserwacji brakujących. W wyniku tej procedury otrzymaliśmy, że nie występują żadne braki w danych. Ponadto łatwo zauważyć, że wszystkie zmienne są odpowiedniego typu.

Dla zmiennych *age*, *tumsize*, *posnodal*, *prm* oraz *esm* w Tabeli 6.30 przedstawiliśmy wartości podstawowych wskaźników statystycznych.

Tabela 6.30: Wartości wskaźników dla zmiennych ciągłych ze zbioru danych `GBSG`

	age	*tumsize*	*posnodal*	*prm*	*esm*
Średnia	53.05	29.33	5.01	110.00	96.25
Wariancja	102.43	204.38	29.98	40938.06	23434.70
Odchylenie standardowe	10.12	14.30	5.48	202.33	153.08
Mediana	53.00	25.00	3.00	32.50	36.00
Pierwszy kwartyl	46.00	20.00	1.00	7.00	8.00
Trzeci kwartyl	61.00	35.00	7.00	131.75	114.00
Minimum	21.00	3.00	1.00	0.00	0.00
Maksimum	80.00	120.00	51.00	2380.00	1144.00

Biorąc pod uwagę średnie wyniki, pacjentki objęte badaniem były w wieku niewiele ponad 53 lata, rozmiar zdiagnozowanego u nich guza był równy $29.33mm$, miały zajęte około 5 węzłów chłonnych oraz liczność receptorów progesteronowych i estrogenowych równą odpowiednio 110 oraz 96.25 femtomoli. Dla każdej ze zmiennych obserwuje się dość duże zróżnicowanie wyników. Wartości median są zbliżone do wartości średnich z wyjątkiem wyników osiąganych dla zmiennych związanych z licznością obu rodzajów receptorów, gdzie mediana jest zdecydowanie mniejsza od średniej. 25% pacjentek miało mniej niż 46 lat lub było w tym wieku, rozmiar guza mniejszy bądź równy $20mm$, zajęty jeden węzeł chłonny oraz liczność receptorów progesteronowych i estrogenowych mniejszą bądź równą odpowiednio 7 i 8 femtomoli. Podobnie 75% pacjentek nie osiągnęło wieku

61 lat bądź ukończyła ten rok życia, miało rozmiar guza mniejszy bądź równy $35mm$, zajętych 7 bądź mniej węzłów chłonnych oraz liczność receptorów progesteronowych oraz estrogenowych mniejszą bądź równą odpowiednio 131.75 oraz 114 femtomoli. Co więcej, przedziały do których należą wartości przyjmowane przez zmienne sugerują dużą rozpiętość wyników dla każdej z nich.

W rozpatrywanym zbiorze nie wszystkie zmienne są ciągłe. Z tego powodu dla zmiennych dyskretnych stworzyliśmy tabele liczności.

Tabela 6.31: Tabela liczności dla zmiennej *htreat* ze zbioru danych `GBSG`

Poziom	Liczba obserwacji
0	440
1	246

Wyniki dotyczące zmiennej *htreat* zostały przedstawione w Tabeli 6.31. W analizowanym zbiorze danych znajduje się więcej informacji na temat pacjentek, które nie były poddane terapii hormonalnej. Nie otrzymało jej bowiem 64.14% z nich.

Tabela 6.32: Tabela liczności dla zmiennej *menostat* ze zbioru danych `GBSG`

Poziom	Liczba obserwacji
1	290
2	396

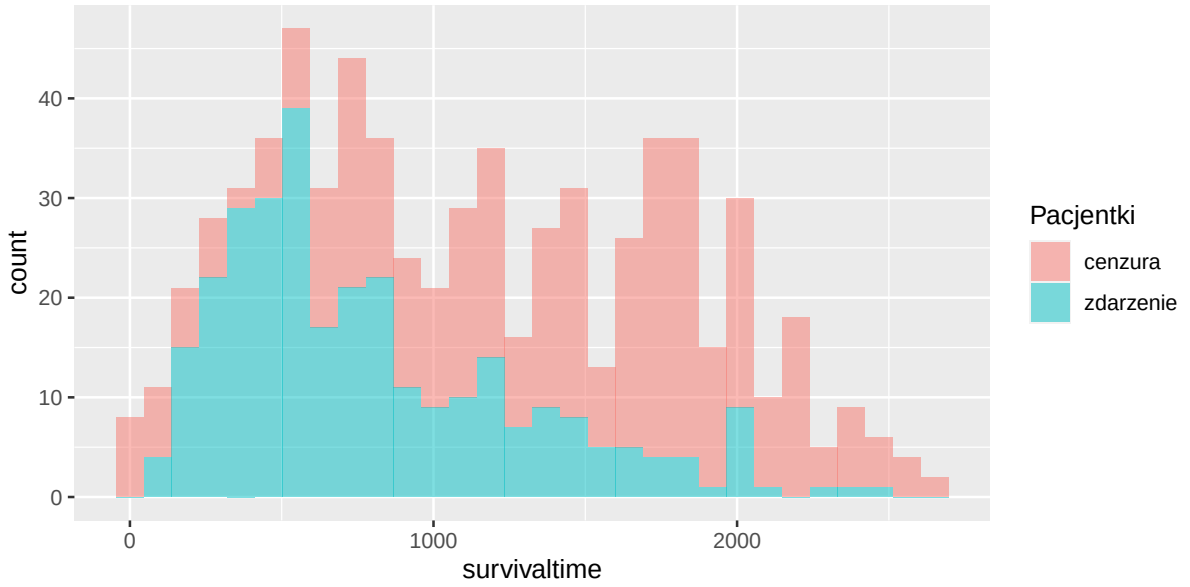
Zgodnie z wynikami przedstawionymi w Tabeli 6.32, zdecydowana większość pacjentek, bo aż 396 z nich znajdowała się w okresie pomenopauzalnym. Stanowią one 57.73% wszystkich obserwacji. Prawdopodobnie jest to bezpośrednio związane ze średnim wiekiem osiąganym przez pacjentki. Jak już wcześniej bowiem wspomniano średnia wieku badanych wynosi ponad 50 lat.

Tabela 6.33: Tabela liczności dla zmiennej *tumgrad* ze zbioru danych `GBSG`

Poziom	Liczba obserwacji
1	81
2	444
3	161

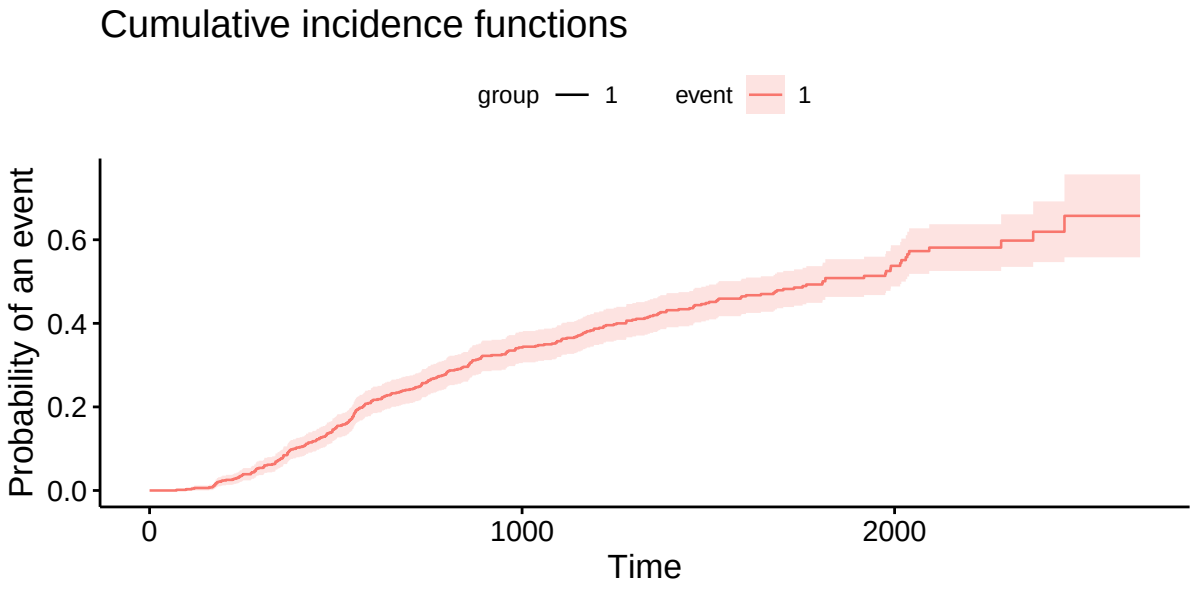
Wyniki przedstawione w Tabeli 6.33 pozwalają stwierdzić, że stopień zaawansowania nowotworu również nie był rozłożony równomiernie w obrębie trzech rozpatrywanych grup. Największa grupa pacjentek stanowiąca frakcję 64.72% chorowała na nowotwór w drugim stopniu zaawansowania. W najbardziej zaawansowanym stadium choroby znajdowało się z kolei niemal dwa razy więcej pacjentek niż w jej początkowym etapie.

Histogram czasu przeżycia z podziałem na dane cenzurowane i kompletne jest widoczny na Rysunku 6.31.



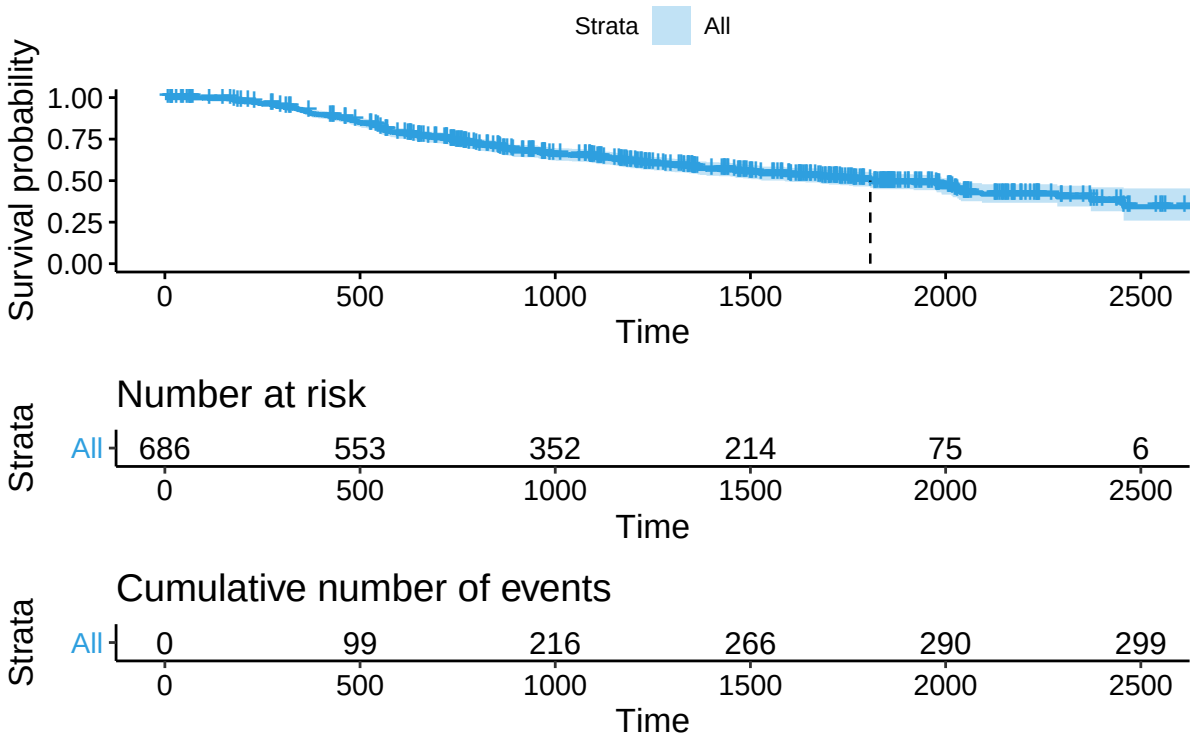
Rysunek 6.31: Histogram czasu przeżycia dla zbioru danych `GBSG`

Wyniki, które zaprezentowaliśmy na Rysunku 6.31 pozwalają wysnuć wniosek, że rozkład danych cenzurowanych i kompletnych jest w przybliżeniu podobny. Wraz ze wzrostem czasu przeżycia obserwujemy bowiem zmiejszenie liczby obserwacji – najwięcej danych jest związanych z krótszym czasem przeżycia. Warto dodać, że w zbiorze danych znajduje się 387 obserwacji cenzurowanych, co stanowi frakcję 56.41%.



Rysunek 6.32: Skumulowana częstość występowania zdarzenia dla zbioru danych `GBSG`

Prawdopodobieństwo wystąpienia zdarzenia przed danym momentem czasowym przedstawione na Rysunku 6.32 oczywiście wzrasta wraz z upływem czasu, lecz wzrost ten nie jest dynamiczny. Po 2000 dniach, czyli po prawie 5.5 roku skumulowana częstość wystąpienia zdarzenia wynosi około 0.6.



Rysunek 6.33: Estymowana funkcja przeżycia wraz z medianą czasu przeżycia dla zbioru danych `GBSG`

Biorąc pod uwagę wyniki wcześniejszych analiz nie powinno spodziewać się, że funkcja przeżycia będzie szybko dążyć do zera. Intuicje te potwierdza Rysunek 6.33. Dopiero po 2500 dniach, czyli około 7 latach mała liczba pacjentek jest zagrożonych wystąpieniem zdarzenia. Prawdopodobieństwo remisji jest dość wysokie przez cały okres obserwacji.

W następnym kroku możemy przejść do dopasowania modelu ryzyka addytywnego. W tym celu skorzystano z funkcji `ahaz()` z pakietu `ahaz`. Najpierw jednak należy odpowiednio przygotować dane, bowiem przy natychmiastowej próbie ich wykorzystania pojawiają się pewne problemy. Pierwszy z nich dotyczy braku akceptacji zmiennych sfaktoryzowanych. Z tego powodu odpowiednie zmienne należy skonwertować do typu numerycznego.

```
GBSG$htreat<-as.numeric(GBSG$htreat)
GBSG$menostat<-as.numeric(GBSG$menostat)
GBSG$tumgrad<-as.numeric(GBSG$tumgrad)
```

Dodatkowym problemem jest wymóg braku powtarzających się wartości czasów przeżycia. W dokumentacji zaproponowano jednak gotową strategię postępowania w takiej sytuacji. Mianowicie do każdego z czasów przeżycia należy dodać odpowiednio pomniejszoną wartość wylosowaną z rozkładu jednostajnego. Dla zapewnienia powtarzalności otrzymywanych wyników wcześniej ustawiliśmy ziarno przy pomocy funkcji `set.seed()`.

```
set.seed(10101)
GBSG$rfst<-GBSG$rfst+runif(nrow(GBSG))*1e-2
```

Po tych modyfikacjach możliwe stało się dopasowanie modelu postaci

```
add.haz.full<-ahaz(Surv(GBSG$rfst,GBSG$cens),predictors)
```

gdzie zmienna `predictors` jest podzbiorem zbioru danych zawierającym wszystkie predyktory.

Tabela 6.34: Współczynniki wraz z wybranymi wartościami dla pełnego modelu addytywnych hazardów

	β	SE_β	Z	$P(>\|z\|)$
htreat	-1.36059e-04	0.00005	-2.92364	0.00346
age	-3.20770e-06	0.00000	-0.84508	0.39807
menostat	1.02397e-04	0.00007	1.54850	0.12150
tumsize	2.79218e-06	0.00000	1.40015	0.16147
tumgrad	1.13899e-04	0.00004	3.16505	0.00155
posnodal	3.84541e-05	0.00001	5.57964	0.00000
prm	-4.39252e-07	0.00000	-4.95638	0.00000
esm	4.85434e-08	0.00000	0.26362	0.79207

Z wyników zaprezentowanych w Tabeli 6.34 wynika, że nie wszystkie zmienne wchodzące w skład modelu są związane z wartością p mniejszą niż 0.05. Dla zmiennych *age*, *menostat*, *tumsize* oraz *esm* nie ma bowiem podstaw do odrzucenia hipotezy o nieistotności współczynnika. Ponadto wartości otrzymanych błędów standardowych nie są duże. Funkcja `summary()` zwraca także wynik testu Walda, który weryfikuje hipotezę zerową zakładającą, że model jest nieistotny. Otrzymując wartość p równą $1.11e-16$ należy zatem odrzucić tę hipotezę, co oznacza, że model ma dość dobre własności. Jednak z powodu nieistotności wcześniej wymienionych zmiennych zastosujemy metody krokowe wyboru zmiennych do modelu, które pozwolą na jego lepsze dopasowanie.

Wybór zmiennych do modelu wykonujemy na dwa sposoby. Pierwszy z nich wykorzystuje wbudowaną funkcję `ahazisis()` z pakietu `ahaz`, natomiast drugi jest ręcznie wykonaną eliminacją wsteczną. Druga ze wspomnianych procedur będzie polegać na eliminacji z pełnego modelu kolejno zmiennych, których wartości p w teście istotności współczynnika są największe i jednocześnie większe od wcześniej ustalonego poziomu istotności $\alpha = 0.05$.

W przypadku procedury wykorzystującej wbudowaną funkcję istnieje możliwość wyboru funkcji kontroli dla regularyzacji. Domyślna metoda korzystająca z naturalnej funkcji straty i walidacji krzyżowej w naszej sytuacji daje niestabilne wyniki. Dlatego też zdecydowaliśmy o wyborze zmiennych do modelu przy pomocy funkcji kontroli dla regularyzacji inspirowanej podejściem bayesowskiego kryterium informacyjnego BIC. Ranking będzie przeprowadzany w oparciu o wartość statystyki $|Z|$.

```
step<-ahazisis(Surv(GBSG$rfst,GBSG$cens),predictors,
                  rank="z",tune=bic.control())
```

Następnie należy sprawdzić jakie predyktory wybrał użyty algorytm. Funkcja `ahazisis()` zwraca indeksy tychże obiektów.

```
step$ISISind

## [1] 1 5 6 7
```

Otrzymaliśmy zatem, że zmienne jakie powinno się wybrać do modelu to *htreat*, *tumgrad*, *posnodal* oraz *prm*.

W przypadku wcześniej opisanej eliminacji wstecznej opierającej się na wartości p w teście istotności współczynnika należało usunąć kolejno zmienne *esm*, *age*, *menostat* oraz *tumsize*. Otrzymany zestaw zmiennych, który należy wykorzystać do utworzenia modelu jest zatem identyczny jak w przypadku wykorzystania wbudowanej funkcji `ahazisis()`. Poniżej zbudowaliśmy model dla tychże predyktorów.

```
step.add.haz<-ahaz(Surv(GBSG$rfst,GBSG$cens),selected.predictors)
```

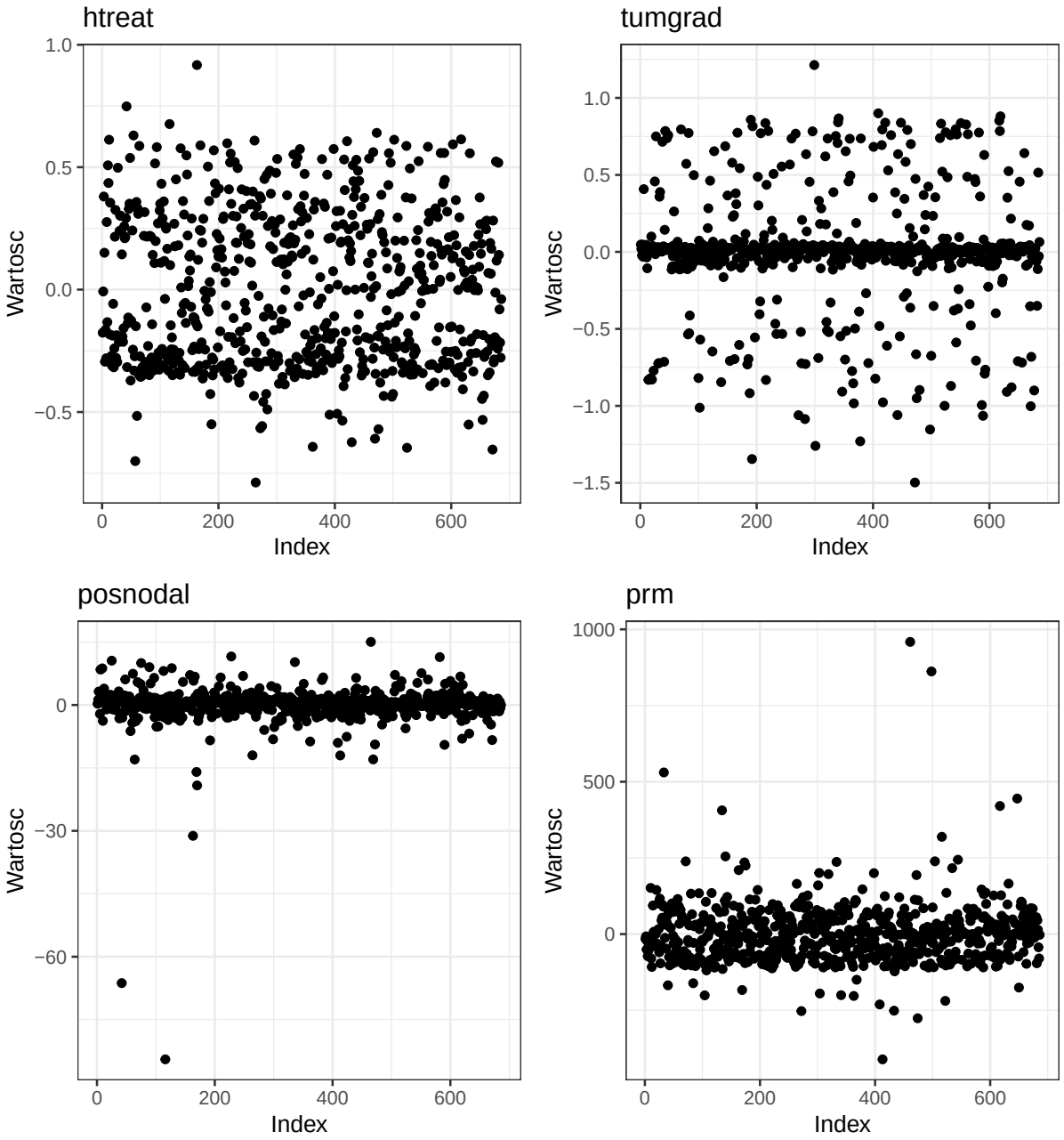
Podobnie jak poprzednio przyjrzymy się wartościom otrzymanym po zastosowaniu funkcji `summary()`.

Tabela 6.35: Współczynniki wraz z wybranymi wartościami dla wybranego modelu addytywnych hazardów

	β	SE_β	Z	$P(>\|z\|)$
htreat	-1.24393e-04	0.00004	-2.79074	0.00526
tumgrad	1.18146e-04	0.00004	3.29420	0.00099
posnodal	4.09295e-05	0.00001	6.14944	0.00000
prm	-4.31406e-07	0.00000	-5.76124	0.00000

Wyniki zamieszczone w Tabeli 6.35 pozwalają stwierdzić, że dla każdej ze zmiennych należy odrzucić hipotezę zerową. Podobnie jak poprzednio zakłada ona, że odpowiedni współczynnik jest równy zero. Ponadto wnioskujemy, że dla rozpatrywanego modelu należy odrzucić hipotezę zerową o jego nieistotności, ponieważ wartość p otrzymana w teście Walda wynosi 0. Wartości błędów standardowych dla poszczególnych współczynników są nadal bardzo małe. Co więcej, w przypadku zmiennej *htreat* obserwujemy nieznaczne jego zmniejszenie. Przechodząc do interpretacji współczynników modelu widzimy, że różnica hazardów w grupach leczonych i nieleczonych terapią hormonalną jest równa $-1.24393e-04$, co oznacza spadek ryzyka nawrotu przy zastosowaniu terapii hormonalnej i ustalonych wartościach pozostałych predyktorów o wartość bezwzględną tego współczynnika. Wzrost stopnia zaawansowania choroby o jeden przy ustalonych wartościach pozostałych zmiennych objaśniających powoduje wzrost ryzyka nawrotu nowotworu o $1.18146e-04$. Z kolei zwiększenie liczby zajętych węzłów chłonnych o jeden przy ustalonych wartościach pozostałych ze zmiennych objaśniających powoduje wzrost wartości ryzyka nawrotu choroby o $4.09295e-05$. Wartość ostatniego ze współczynników sugeruje, że wzrost liczności receptorów progesteronowych o 1000 femtomoli przy ustalonych wartościach pozostałych zmiennych zmniejsza ryzyko wystąpienia zdarzenia o $4.31406e-04$.

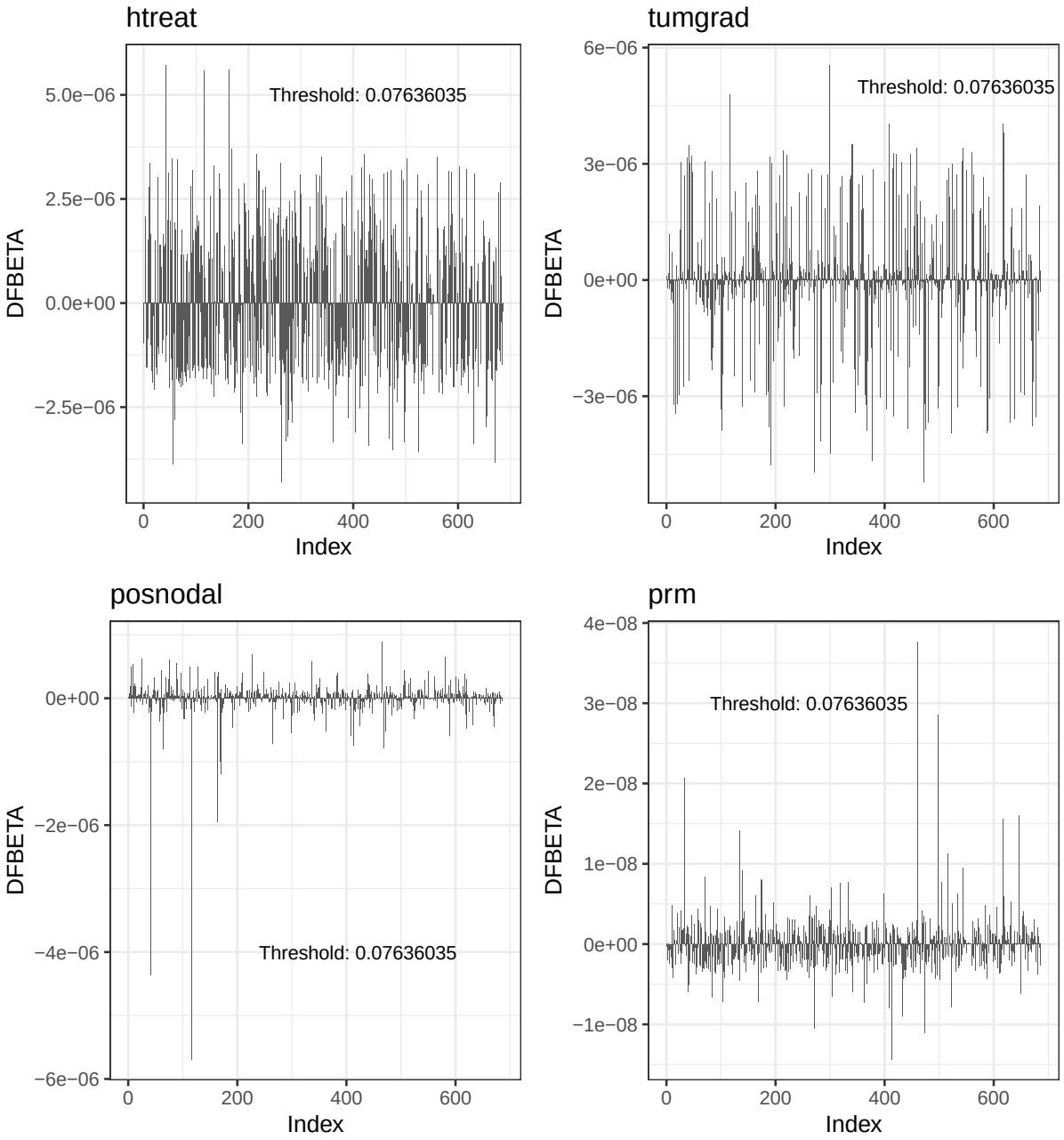
Ważnym elementem analizy modelu ryzyka addytywnego jest wizualizacja scałkowanych reszt matryngałowych (ang. *integrated martingale residuals*). Rezydua te dla każdej ze zmiennych wchodzących w skład modelu zostały przedstawione na Rysunku 6.34.



Rysunek 6.34: Scałkowane reszty martyngałowe dla wybranego modelu ryzyka addytywnego

Scałkowane rezydua matryngałowe są niezależnymi zmiennymi losowymi o jednakowym rozkładzie, a ich rozkład asymptotyczny powinien być wielowymiarowym rozkładem normalnym o średniej zero. Otrzymane wartości średnie reszt dla zmiennych *htreat*, *tumgrad*, *posnodal* oraz *prm* wynoszą odpowiednio $5.577784e-15$, $4.046481e-15$, $9.048826e-16$ oraz $-1.032565e-13$, a zatem zgodnie z założeniami są równe w przybliżeniu zero.

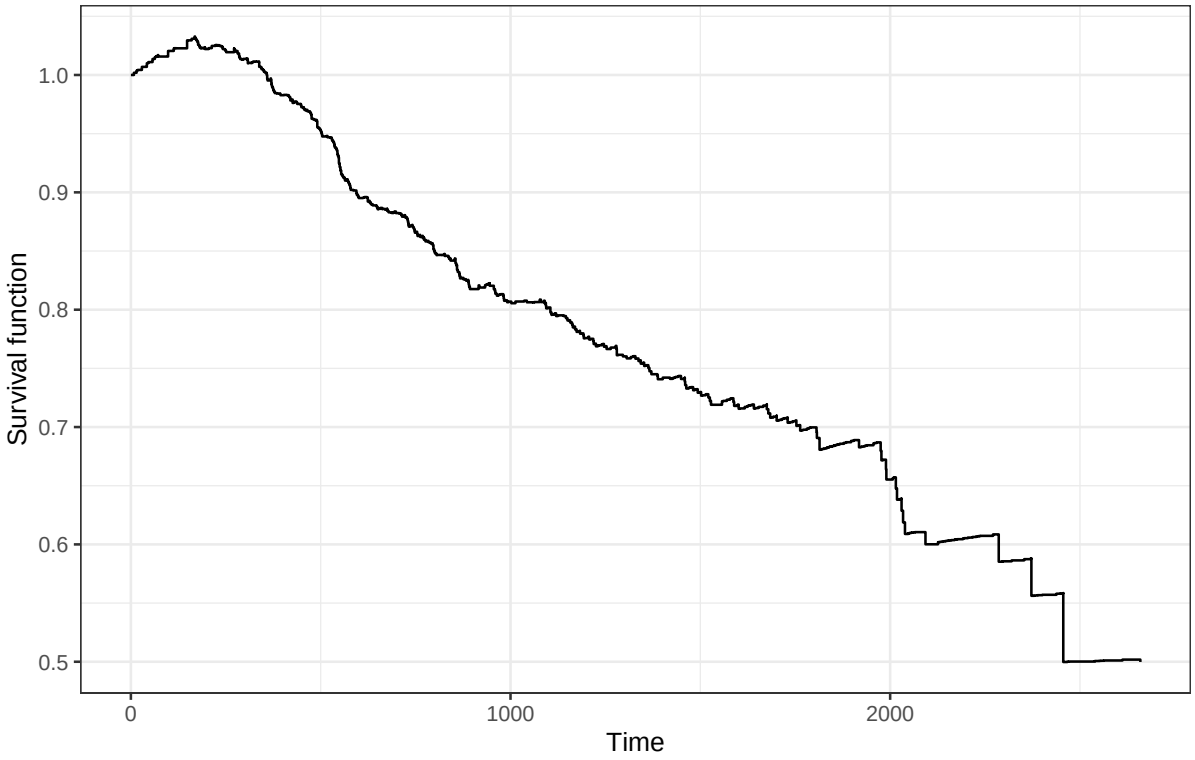
W ostatnim kroku tak jak w poprzednich podrozdziałach wykonamy diagnostykę obserwacji wpływowych przy wykorzystaniu wartości $DFBETA$. Otrzymane wyniki są widoczne na Rysunku 6.35.



Rysunek 6.35: Wartości $DFBETA$ dla wybranego modelu ryzyka addytywnego

W przypadku danych rozpatrywanych w niniejszym podrozdziale próg klasyfikacji jako wpływowej jest równy 0.07636035. Oznacza to, że w wybranym modelu nie obserwuje się żadnych obserwacji nietypowych. Zgodnie z wynikami przedstawionymi na Rysunku 6.35 są wartości $DFBETA$ dla każdej ze zmiennych są bowiem bardzo małe.

Na koniec przedstawimy estymowaną funkcję przeżycia dla wybranego modelu ryzyka addytywnego. Wyniki przedstawiliśmy na Rysunku 6.36.



Rysunek 6.36: Estymowana funkcja przeżycia dla wybranego modelu ryzyka addytywnego

Rysunek 6.36 przedstawia estymowaną funkcję przeżycia przez model ryzyka addytywnego. Została ona otrzymana na podstawie wartości `cumhaz`, które zgodnie z dokumentacją biblioteki `ahaz` powinny być utożsamiane z wartościami hazardu skumulowanego. Niestety niektóre ze zwracanych wartości są ujemne co stoi w sprzeczności z własnościami hazardu skumulowanego. W rezultacie estymowana funkcja przeżycia początkowo przyjmuje wartości większe od 1 co również zdecydowanie nie powinno mieć miejsca. Budzi to więc wątpliwości na temat tego, czy funkcja została zaimplementowana poprawnie przez autorów wykorzystanej biblioteki.

Podsumowanie

Celem niniejszej pracy była prezentacja wybranych modeli nieproporcjonalnych hazardów, które mogą być zastosowane w przypadku, gdy założenie proporcjonalności hazardów powszechnie stosowanego modelu Coxa nie jest spełnione. Warto dodać, że model proporcjonalnych hazardów Coxa jest modelem, który bardzo często używany jest w analizie przeżycia, jednak ograniczenia z nim związane sprawiają, że często zachodzi potrzeba wykorzystania modelu, który nie nakłada na nas wcześniej wspomnianych wymogów. Warto jednak zauważyć, że w pewnych sytuacjach model proporcjonalnych hazardów Coxa mimo wszystko sprawdza się znakomicie. Należy jednak ponownie zaznaczyć, że takie sytuacje są rzadkie.

W pracy opisaliśmy modele, które należy rozważyć w sytuacji, gdy hazard jest nieproporcjonalny. Tematyka modeli nieproporcjonalnych hazardów obecnie dynamicznie się rozwija i jest często rozważana przez autorów licznych artykułów. Można zatem sądzić, że liczba rozwiązań, którymi dysponujemy, w przypadku gdy model proporcjonalnych hazardów Coxa nie może zostać wykorzystany jest dość duża. Dodatkowo wiele z przedstawionych modeli można modyfikować, na przykład dopuszczając sytuację, w której zmienne zależą od czasu. Stwarza to kolejne możliwości prognozy czasu przeżycia. Dla każdego modelu przedstawiliśmy podstawowe informacje na jego temat oraz idee estymacji jego współczynników. Z powodu złożoności omówionych zagadnień, w pracy zaprezentowaliśmy jedynie zarys podjętej problematyki. Bardziej szczegółowy opis modeli i metod estymacji wartości z nimi związanych można znaleźć w artykułach zamieszczonych w Bibliografii.

W pracy samodzielnie zaproponowaliśmy również pomysł modyfikacji przedstawionych modeli w sytuacji, gdy w danych znajduje się duża liczba zer. Główna modyfikacja polegała na wprowadzeniu semiciągłej zmiennej odpowiedzi. Warto zaznaczyć, że problem estymacji współczynników modeli po takiej modyfikacji pozostaje otwarty. Sądząc po liczbie wydanych artykułów, tematyka takich danych w kontekście opisanych modeli nie jest obecnie zbyt popularna, lecz z pewnością znajduje szerokie zastosowanie w praktyce.

W części praktycznej pracy przedstawiliśmy analizę danych rzeczywistych z wykorzystaniem wybranych modeli opisanych w pracy. Została ona przeprowadzona w języku programowania *R*; język *Python* bowiem nie dysponuje dobrymi bibliotekami do tego typu analiz. Modele wybrane przez nas do analizy są modelami najbardziej popularnymi w analizie przeżycia. Ich dopasowanie za każdym razem zostało poprzedzone opisem wybranych danych, co pozwala na dokładniejszą analizę. Przy wyszukiwaniu odpowiednich bibliotek, które są wykorzystywane w pakiecie *R* do dopasowania odpowiednich modeli, można zauważyć, że dla większości z nich możemy znaleźć wiele zaimplementowanych rozwiązań. Jednocześnie niektóre z modeli do chwili obecnej nie są możliwe do zbudowania bezpośrednio przy użyciu gotowych funkcji, ponieważ nie zostały jeszcze dla nich utworzone odpowiednie pakiety.

Bibliografia

[1] D. R. Cox: *Regression Models and Life-Tables*, Journal of the Royal Statistical Society, Series B, Vol. 34, No. 2, pp. 187-220 (1972),

[2] L. Tian, D. Zucker, L. J. Wei: *On the Cox Model With Time-Varying Regression Coefficients*, Journal of the American Statistical Association (2005),

[3] G. Diao, D. Zeng: *Efficient Semiparametric Estimation of Short-Term and Long-Term Hazard Ratios with Right-Censored Data*, Biometrics 69, pp. 840–849 (2013),

[4] A. Y. C. Kuk: *A Non-Proportional Hazards Model with Hazard Ratio Functions Free from Covariate Values*, International Statistical Review (2020),

[5] D. R. Cox: *The Regression Analysis of Binary Sequences*, Journal of the Royal Statistical Society, Vol. 20, No. 2, pp. 215-242 (1958),

[6] V. T. Farewell, R. L. Prentice: *The approximation of partial likelihood with emphasis on case-control studies*, Biometrika, 67, 2, pp. 273-8 (1980),

[7] S. Halabi, S. Dutta, Y. Wu, A. Liu: *Score and Deviance Residuals Based on the Full Likelihood Approach in Survival Analysis*, Pharm Stat. (2020),

[8] F. Xia, J. Ning, X. Huang: *Empirical Comparison of the Breslow Estimator and the Kalbfleisch Prentice Estimator for Survival Functions*, J Biom Biostat. (2018),

[9] A. Tsiatis, *A Large Sample Study of the Estimates for the Integrated Hazard Function in Cox's Regression Model for Survival Data*, Annals of Statistics 9: 93–108 (1981),

[10] P.K. Andersen, O. Borgan, R. D. Gill, N. Keiding, *Statistical Models based on Counting Processes*, Springer Verlag (1993),

[11] R. Braekers, Y. Grouwels: *A semi-parametric cox's regression model for zero-inflated left-censored time to event data*, Communications in Statistics - Theory and Methods, 45:7, 1969-1988 (2016),

[12] P. K. Andersen, M. P. Perme, H. C. van Houwelingen, R. J. Cook, P. Joly, T. Martinussen, J. M. G. Taylor, M. Abrahamowicz, T. M. Therneau: *Analysis of time-to-event for observational studies: Guidance to the use of intensity models*, Statistics in Medicine, Wiley (2020),

[13] S. Guo, D. Zeng: *An overview of semiparametric models in survival analysis*, Journal of Statistical Planning and Inference, Elsevier (2013),

[14] S. Kirmani, R. C. Gupta: *On the proportional odds model in survival analysis*, The Institute of Statistical Mathematics (2001),

[15] P. McCullagh: *On the Elimination of Nuisance Parameters in the Proportional Odds Model*, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 46, No. 2, pp. 250-256, Wiley for the Royal Statistical Society (1984),

[16] T. Martinussen, T. H. Scheike: *Dynamic Regression Models for Survival Data*, Statistics for Biology and Health, Springer (2006),

[17] W. Lu, A. A. Tsiatis: *Semiparametric transformation models for the case-cohort study*, Biometrika, pp. 207–214 (2006),

[18] D. Zeng, D. Y. Lin: *Efficient estimation of semiparametric transformation models for counting processes*, Biometrika, pp. 627–640 (2006),

[19] K. Chen, Z. Jin, Z. Ying: *Semiparametric analysis of transformation models with censored data*, Biometrika, pp. 659–668 (2002),

[20] D. M. Dąbrowska, K. A. Doksum: *Partial Likelihood in Transfomation Models with Censored Data*, Scandinavian Journal of Statistics, Vol. 15, Wiley on behalf of Board of the Foundation of the Scandinavian Journal of Statistics, pp. 1-23 (1988),

[21] D. Rava, R. Xu: *Explained Variation under the additive hazards model*, Statistics in Medicine (2020),

[22] L. Chen, F. Gao, C. Xiong, J. P. Miller: *Power and Sample Size Calculations with the Additive Hazards Model*, Journal of Data Science (2012),

[23] X. Xie, H. D. Strickler, X. Xue: *Additive Hazard Regression Models: An Application to the Natural History of Human Papillomavirus*, Hindawi Publishing Corporation, Computational and Mathematical Methods in Medicine (2012),

[24] E. E. Álvarez, J. Ferrario: *Robust Differentiable Functionals for the Additive Hazards Model*, Open Journal of Statistics (2015),

[25] D. Y. Lin, Z. Ying: *Additive Hazards Regression Models for Survival Data*, Proceedings of the First Seattle Symposium in Biostatistics, pp. 185-198 (1997),

[26] Y. Tseng, F. Hsieh, J. Wang: *Joint modelling of accelerated failure time and longitudinal data*, Biometrika Trust, pp. 587–603 (2005),

[27] L. Tian, T. Cai: *On the Accelerated Failure Time Model for Current Status and Interval Censored Data*, Harvard University Biostatistics Working Paper Series (2004),

[28] R. Saikia, M. P. Barman: *A Review on Accelerated Failure Time Models*, International Journal of Statistics and Systems, pp. 311-322 (2017),

[29] X. Yang, M. Abdel-Aty, M. Huan, Y. Peng, Z. Gao: *An accelerated failure time model for investigating pedestrian crossing behavior and waiting times at signalized intersections*, Accident Analysis and Prevention (2015),

[30] L. Kong, J. Cai: *Case-Cohort Analysis with Accelerated Failure Time Model*, Biometrics, Vol. 65, pp. 135-142 (2009),

[31] A. Agresti: *Foundations of Linear and Generalized Linear Models*, Wiley Series in Probability and Statistics (2015),

[32] Y. Xue, E. D. Schifano: *Diagnostics for the Cox model*, Communications for Statistical Applications and Methods, Korean Statistical Society (2017),

[33] A. Winnett, P. Sasieni: *A note on scaled residuals on the proportional hazards model*, Biometrika, pp. 565-571 (2001),

[34] T. Therneau, E. Atkinson: *Concordance* (2020),

[35] T. A. Gerds, M. W, Kattan, M. Schumacher, C. Yu: *Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring*, Statistics in Medicine (2012),

[36] X. Li, S. Deng, L. Li, Y. Jiang: *Outlier Detection Based on Robust Mahalanobis Distance and Its Application*, Open Journal of Statistics, Vol.9, No.1 (2019),

[37] A. Fitrianto, R. L. T. Jiin: *Several Types of Residuals in Cox Regression*, Int. Journal of Math. Analysis, Vol. 7 (2013),

[38] E. Cabana, R. E. Lillo, H. Laniado: *Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators* (2019),

[39] P. M. Grambsch, T. M. Therneau: *Proportional Hazards Tests and Diagnostics Based on Weighted Residuals*, Biometrika, Vol. 81, No. 3, pp. 515-526 (1994),

[40] K. H. Zou, K. Tuncali, S. G. Silverman: *Correlation and Simple Linear Regression*, Statistical Concepts Series (2003),

[41] Z. Zhang, J. Reinikainen, K. A. Adeleke, M. E. Pieterse, C. G. M. Groothuis-Oudshoorn: *Time-varying covariates and coefficients in Cox regression models*, Big-data Clinical Trial Column (2018).