

Raport 2

Analiza Przeżycia

Dominik Kowalczyk 282229 i Matylda Mordal 282240

2026-02-03

Spis treści

1	Lista 9	3
1.1	Zadanie 1	3
1.2	Zadanie 2	4
1.3	Zadanie 3	5
1.4	Zadanie 4	6
1.5	Zadanie 5	6
1.6	Zadanie 6	8
1.7	Zadanie 7	8
1.8	Zadanie 8	11
1.9	Zadanie 9	12
2	Lista 10	14
2.1	Zadanie 1	14
2.2	Zadanie 2	15
2.3	Zadanie 3	15
2.4	Zadanie 4	18
2.5	Zadanie 5	21
2.6	Zadanie 6	24
3	Lista 11	26
3.1	Zadanie 1	26
3.2	Zadanie 2	27
3.3	Zadanie 3	28
3.4	Zadanie 4	30
3.5	Zadanie 5	33
3.6	Zadanie 6	34
4	Lista 12	35
4.1	Zadanie 1	35
4.2	Zadanie 2	37

4.3	Zadanie 3	39
4.4	Zadanie 4	44

1 Lista 9

1.1 Zadanie 1

Zadanie 1

Celem zadania jest estymacja parametrów modelu przyspieszonego czasu awarii (AFT) dla danych *lung*, zakładając rozkład Weibulla czasu przeżycia.

W modelu AFT zakładamy, że funkcja przeżycia $S(x|z)$ dla jednostki o wektorze charakterystyk z wyraża się wzorem:

$$S(x|z) = S_0(\exp(\beta^T z)x),$$

gdzie $S_0(x)$ jest bazową funkcją przeżycia.

Natomiast funkcja hazardu $h(x|z)$ jest postaci:

$$h(x|z) = h_0(\exp(\beta^T z)x) \exp(\beta^T z),$$

gdzie h_0 jest bazową funkcją hazardu.

Bazowe funkcje przeżycia i hazardu przyjmują postać:

$$S_0(x) = \exp(-\lambda_0 x^{\alpha_0}), \quad \text{dla } x > 0,$$

$$h_0(x) = \lambda_0 \alpha_0 x^{\alpha_0-1}, \quad \text{dla } x > 0.$$

Parametry modelu to $\alpha_0 > 0$ i $\lambda_0 > 0$.

Do analizy włączono następujące charakterystyki zmienne:

- *age*: wiek w latach,
- *sex*: płeć (1-mężczyzna, 2-kobieta),
- *ph.ecog*: skala sprawności ECOG wg. lekarza (0-sprawność prawidłowa, 5-zgon),
- *ph.karno*: skala sprawności Karnofsky'ego wg. lekarza (100-sprawność prawidłowa, 0-zgon).

Kod

Przygotowanie danych

```
lung <- survival::lung
dane <- lung[, c("time", "status", "age", "sex", "ph.ecog", "ph.karno")]
#Usuwamy braki
dane <- na.omit(dane)
#Centrujemy
age_m <- mean(dane$age)
ph_karno_m <- mean(dane$ph.karno)
```

```
dane$age_cent <- dane$age - age_m
dane$ph_karno_cent <- dane$ph.karno - ph_karno_m
```

W przedstawionym kodzie, z wykorzystaniem funkcji *survreg* z pakietu *survival* dopasowano parametryczny model przyspieszonego czasu awarii (AFT) z rozkładem Weibulla do danych. Na podstawie wyników modelu obliczono parametry.

```
model_AFT <- survreg(Surv(time, status) ~
  age_cent + as.factor(sex) + as.factor(ph.ecog) + ph_karno_cent,
  data = dane, dist = "weibull")

#Parametry
beta_AFT <- -summary(model_AFT)$coefficients
mu_AFT <- model_AFT$coefficients[1]
sigma_AFT <- model_AFT$scale
alpha_AFT <- 1/sigma_AFT
lambda_AFT <- exp(-mu_AFT * alpha_AFT)
```

Tabela 1: Oszacowane parametry modelu przyspieszonego czasu awarii

Zmienna	β
(Intercept)	-6.3011805
age_cent	0.0086018
as.factor(sex)2	-0.4082801
as.factor(ph.ecog)1	0.4215608
as.factor(ph.ecog)2	0.9261135
as.factor(ph.ecog)3	1.6872434
ph_karno_cent	0.0100567

1.2 Zadanie 2

Oszacowane wartości parametrów przedstawiono w tabeli 1. W tym modelu dodatnie wartości współczynników oznaczają skracanie czasu życia, natomiast ujemne jego wydłużanie.

Wyraz wolny *Intercept* ma wartość ujemną i ma głównie znaczenie techniczne. Odpowiada bazowemu poziomowi czasu przeżycia dla pacjenta referencyjnego (przeciętny wiek i ph.karno, mężczyzna, ECOG = 0) i stanowi punkt odniesienia dla interpretacji pozostałych współczynników, sam w sobie nie ma bezpośredniej interpretacji klinicznej.

Wiek pacjenta (*age_cent*) ma dodatni współczynnik, co oznacza, że wraz ze wzrostem wieku ponad średnią czas przeżycia ulega skróceniu. Wpływ ten jest jednak stosunkowo niewielki. Współczynnik dla płci (kobiety) jest ujemny, co wskazuje na wydłużenie czasu przeżycia w porównaniu z mężczyznami. Bardzo wyraźny wpływ na czas przeżycia mają zmienne opisujące sprawność pacjenta według skali ECOG. Wszystkie współczynniki dla kategorii ph.ecog są dodatnie, a ich wartości rosną wraz z pogorszeniem stanu pacjenta. Oznacza to, że im wyższy

poziom ECOG, tym krótszy oczekiwany czas życia. Zmienna *ph.karno_cent* również ma dodatni współczynnik, co oznacza, że niższa sprawność według skali Karnofsky’ego (wartości poniżej średniej) wiąże się ze skróceniem czasu przeżycia.

1.3 Zadanie 3

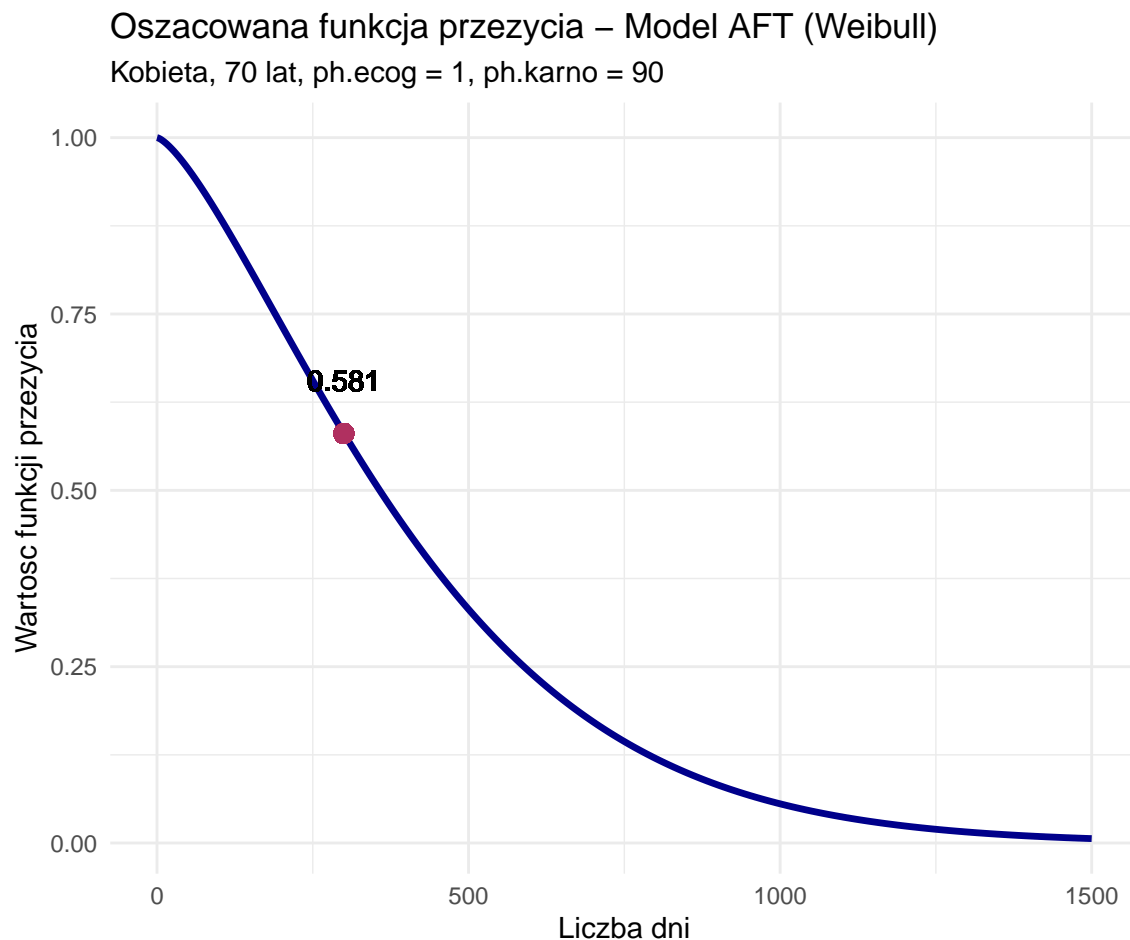
Kod

W kodzie zdefiniowano funkcję *S_AFT*, która oblicza wartość funkcji przeżycia w modelu AFT z rozkładem Weibulla dla zadanego czasu x oraz wektora charakterystyk z . Funkcja ta wykorzystuje wcześniej oszacowane parametry modelu: λ_0 , α_0 oraz β . Następnie zdefiniowano wektor cech z odpowiadający kobiecie w wieku 70 lat o charakterystyce *ph.ecog*=1 i *ph.karno*=90 (po centrowaniu). Na końcu obliczono prawdopodobieństwo przeżycia dłużej niż 300 dni dla tej kombinacji cech.

```
S_AFT <- function(x, z_vec) {  
  exp(-lambda_AFT * exp(alpha_AFT * sum(beta_AFT[-1] * z_vec)) * x^alpha_AFT)  
}  
  
z <- c(70 - age_m, 1, 1, 0, 0, 90 - ph_karno_m)  
  
p_300 <- S_AFT(300, z)
```

Szacowane prawdopodobieństwo przeżycia przez tę konkretną kobietę dłużej niż 300 dni wynosi około 58.06%.

1.4 Zadanie 4



Rysunek 1: Oszacowana funkcja przeżycia z modelu AFT (Weibull) dla kobiety w wieku 70 lat o cechach $ph.ecog = 1$ i $ph.karno = 90$, zaznaczono punkt odpowiadający wartości w $t = 300$.

Wykres 1 przedstawia oszacowaną funkcję przeżycia $S(x|z)$ uzyskaną z modelu AFT (Weibull) dla kobiety w wieku 70 lat, z $ph.ecog=1$ i $ph.karno=90$. Oś pozioma to liczba dni, oś pionowa to wartość funkcji przeżycia (prawdopodobieństwo przeżycia dłużej niż x dni). Krzywa zaczyna blisko 1 i maleje w miarę upływu czasu, co odzwierciedla spadające prawdopodobieństwo przeżycia. Zaznaczony na krzywej różowy punkt odpowiada wynikowi uzyskanemu w zadaniu trzecim i wskazuje, że szacowane prawdopodobieństwo przeżycia przez tę kobietę co najmniej 300 dni wynosi 0.5806. Kształt krzywej jest zgodny z $\alpha_0=1.388>1$ (hazard rosnący), stąd funkcja przeżycia maleje coraz szybciej.

1.5 Zadanie 5

Celem zadania jest estymacja parametrów modelu proporcjonalnych hazardów (PH) dla danych *lung*, przy założeniu rozkładu Weibulla czasu przeżycia.

W modelu PH zakładamy, że funkcja hazardu dla jednostki o wektorze charakterystyk z ma postać:

$$h(x|z) = h_0(x) \exp(\beta^T z),$$

gdzie $h_0(x)$ jest bazową funkcją hazardu, a $\exp(\beta^T z)$ opisuje wpływ zmiennych objaśniających na poziom hazardu.

Odpowiadająca funkcja przeżycia wyraża się wzorem:

$$S(x|z) = [S_0(x)]^{\exp(\beta^T z)},$$

gdzie $S_0(x)$ jest bazową funkcją przeżycia.

Zakładamy, że bazowe funkcje odpowiadają rozkładowi Weibulla, czyli:

$$h_0(x) = \lambda_0 \alpha_0 x^{\alpha_0 - 1}, \quad \text{dla } x > 0,$$

$$S_0(x) = \exp(-\lambda_0 x^{\alpha_0}), \quad \text{dla } x > 0,$$

gdzie parametry modelu $\alpha_0 > 0$, $\lambda_0 > 0$.

Do modelu włączono następujące zmienne objaśniające:

- *age*: wiek w latach,
- *sex*: płeć (1-mężczyzna, 2-kobieta),
- *ph.ecog*: skala sprawności ECOG wg. lekarza (0-sprawność prawidłowa, 5-zgon),
- *ph.karno*: skala sprawności Karnofsky'ego wg. lekarza (100-sprawność prawidłowa, 0-zgon).

Współczynniki β w modelu PH opisują wpływ tych zmiennych na hazard zgonu. Wartość $\exp(\beta_k)$ interpretowana jest jako iloraz hazardów dla jednostek różniących się o jednostkę w k -tej zmiennej, przy pozostałych cechach stałych.

Kod

W poniższym kodzie, przy użyciu funkcji *phreg* z pakietu *eha*, dopasowano parametryczny model proporcjonalnych hazardów (PH) z rozkładem Weibulla. Na podstawie uzyskanych wyników obliczono parametry modelu.

```
model_PH <- phreg( Surv(time, status == 2) ~
  age_cent + as.factor(sex) + as.factor(ph.ecog) + ph.karno_cent,
  data = dane, dist = "weibull" )

beta_PH <- model_PH$coefficients[-c(7, 8)]
mu_PH <- model_PH$coefficients["log(scale)"]
sigma_PH <- exp(model_PH$coefficients["log(shape)"])
lambda_PH <- exp(-mu_PH * sigma_PH)
alpha_PH <- sigma_PH
```

Tabela 2: Oszacowane parametry modelu proporcjonalnych hazardów

Zmienna	β
age_cent	0.0119361
as.factor(sex)2	-0.5665420
as.factor(ph.ecog)1	0.5849707
as.factor(ph.ecog)2	1.2851033
as.factor(ph.ecog)3	2.3412704
ph_karno_cent	0.0139550

1.6 Zadanie 6

W tabeli 2 przedstawiono oszacowane współczynniki parametrycznego modelu proporcjonalnych hazardów. W tym modelu parametry opisują wpływ poszczególnych cech pacjenta na poziom hazardu zgonu, czyli chwilowego ryzyka wystąpienia zdarzenia. Współczynnik dodatni oznacza zwiększenie hazardu, a więc krótszy czas przeżycia, natomiast współczynnik ujemny oznacza zmniejszenie hazardu, czyli dłuższe przeżycie.

Wiek działa niekorzystnie, wraz ze wzrostem wieku powyżej średniej zwiększa się ryzyko zgonu. Oznacza to, że starsi pacjenci mają gorsze rokowanie niż młodsi o podobnych pozostałych charakterystykach.

Płeć istotnie różnicuje przeżycie. Współczynnik dla kobiet ma znak ujemny, co wskazuje na niższe ryzyko zgonu w porównaniu z mężczyznami. Kobiety charakteryzują się więc korzystniejszym rokowaniem przy tych samych pozostałych cechach.

Bardzo silny wpływ na hazard zgonu ma stan sprawności oceniany w skali ECOG. Wraz z pogarszaniem się stanu ogólnego pacjenta ryzyko zgonu wyraźnie rośnie. Oznacza to, że gorsza sprawność jest jednym z najsilniejszych niekorzystnych czynników prognostycznych w analizowanym modelu.

Niewielką, ale dodatnią wartość parametru obserwujemy dla skali Karnofsky’ego. Lepsza sprawność pacjenta wiąże się z niższym ryzykiem zgonu, natomiast gorsza ze wzrostem hazardu.

Podsumowując, model proporcjonalnych hazardów wskazuje, że największe znaczenie dla ryzyka zgonu ma ogólny stan kliniczny pacjenta, mierzony skalami sprawności, następnie płeć, a także wiek, który również działa niekorzystnie, choć jego wpływ jest słabszy niż wpływ miar sprawności.

1.7 Zadanie 7

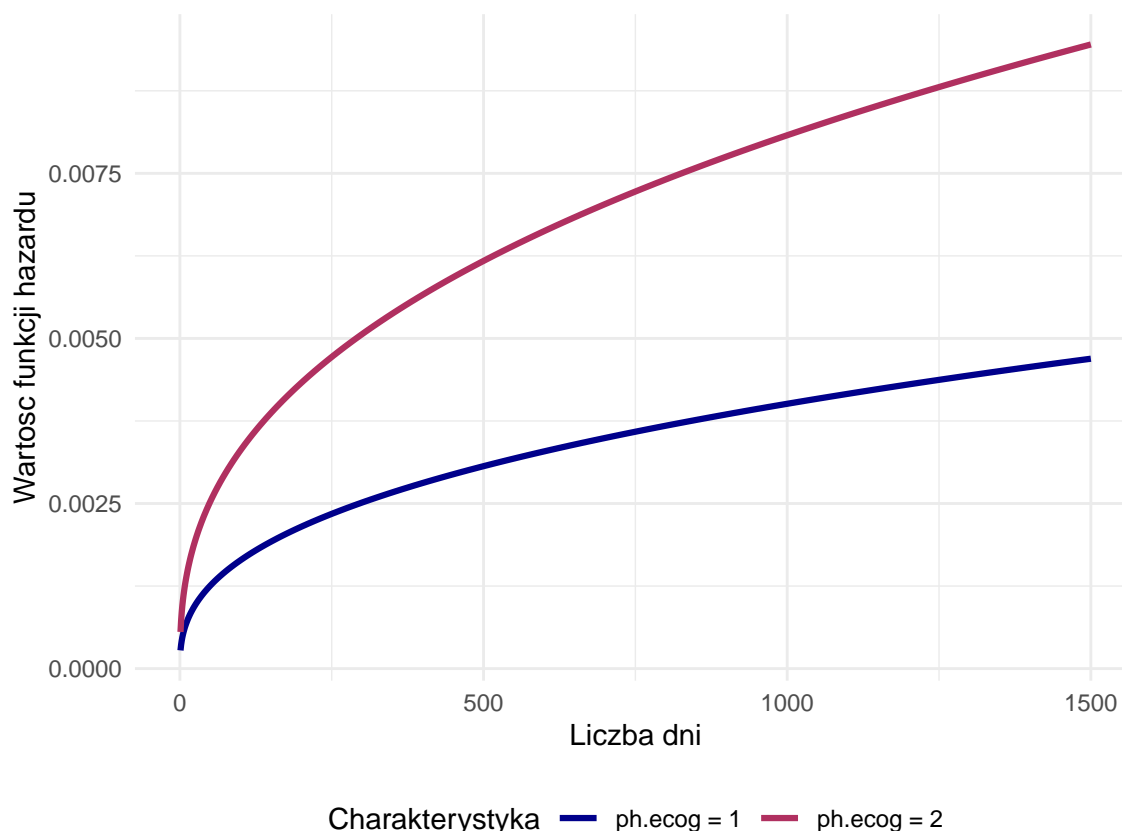
Kod

W poniższym kodzie zdefiniowano funkcję h_PH , która oblicza wartość funkcji hazardu w modelu proporcjonalnych hazardów (PH) z rozkładem Weibulla.


```
h_PH <- function(x, z_vec) {
  lambda_PH * alpha_PH * x^(alpha_PH - 1) * exp(sum(beta_PH * z_vec))
}
```

Oszacowana funkcja hazardu – Model PH (Weibull)

Kobieta, 70 lat, ph.karno = 90

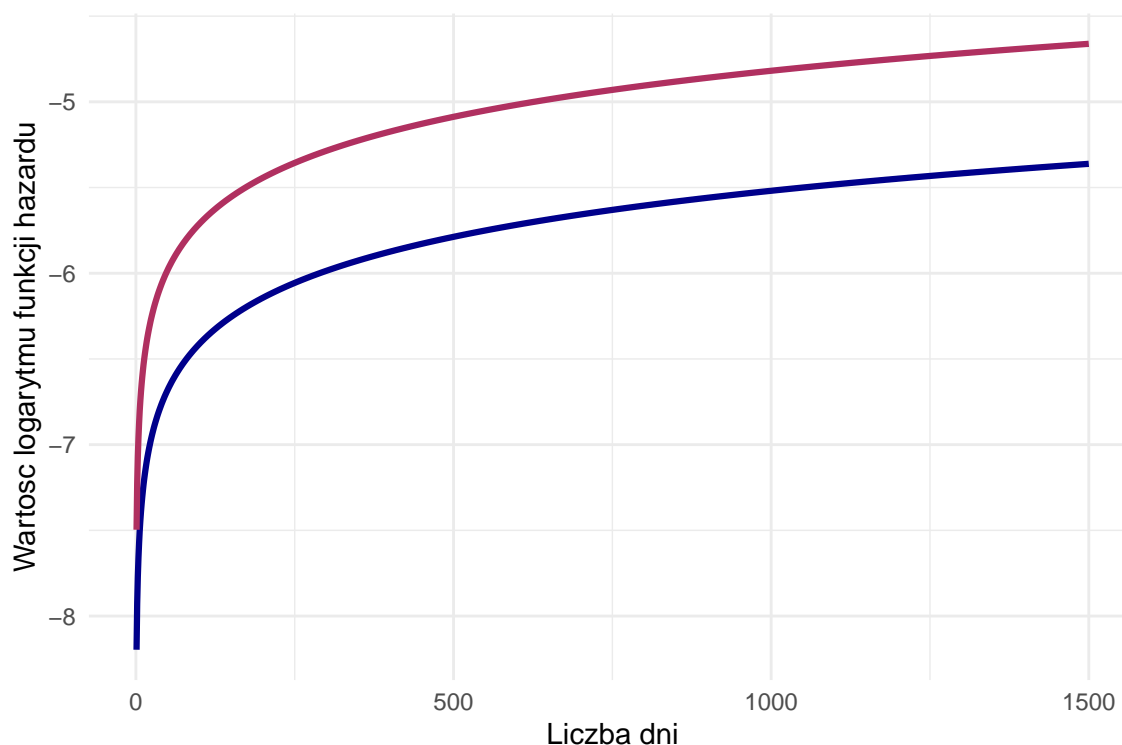


Rysunek 2: Oszacowana funkcja hazardu w modelu PH (Weibull) dla kobiety w wieku 70 lat o $ph.karno = 90$, porównanie dla $ph.ecog = 1$ oraz $ph.ecog = 2$.

Rysunek 2 przedstawia przebieg oszacowanej funkcji hazardu w parametrycznym modelu proporcjonalnych hazardów dla kobiety w wieku 70 lat o wartości $ph.karno = 90$, porównując dwie różne wartości sprawności według skali ECOG. Na wykresie widoczne są dwie krzywe: dla $ph.ecog = 1$ oraz $ph.ecog = 2$. Krzywa odpowiadająca $ph.ecog = 2$ przebiega wyraźnie wyżej niż krzywa dla $ph.ecog = 1$, co oznacza, że pacjentki o gorszej sprawności mają w każdym momencie wyższy hazard, czyli większe chwilowe ryzyko zgonu. Obie funkcje rosną wraz z czasem, co jest zgodne z właściwościami rozkładu Weibulla przy oszacowanym parametrze $\alpha_0 = 1.388 > 1$, hazard zwiększa się wraz z upływem czasu.

Logarytm funkcji hazardu – Model PH (Weibull)

Kobieta, 70 lat, ph.karno = 90



Charakterystyka — $ph.ecog = 1$ — $ph.ecog = 2$

Rysunek 3: Logarytm oszacowanej funkcji hazardu w modelu PH (Weibull) dla kobiety w wieku 70 lat o $ph.karno = 90$, porównanie dla $ph.ecog = 1$ oraz $ph.ecog = 2$.

Na Rysunku przedstawiono logarytm oszacowanej funkcji hazardu w parametrycznym modelu proporcjonalnych hazardów Weibulla dla kobiety w wieku 70 lat i wartości $ph.karno = 90$, porównując dwie wartości sprawności według skali ECOG. Na wykresie widoczne są dwie rosnące krzywe: dla $ph.ecog = 1$ oraz $ph.ecog = 2$. Krzywa odpowiadająca $ph.ecog = 2$ przebiega wyżej na całej długości osi czasu, co oznacza, że pacjentki o gorszej sprawności mają wyższy log-hazard, a więc również wyższe chwilowe ryzyko zgonu niż pacjentki z $ph.ecog = 1$.

Obie krzywe mają zbliżony kształt i pozostają od siebie oddalone o stałą wartość, co jest zgodne z założeniem proporcjonalności hazardów, różnice między grupami są stałe w czasie. Rosnący charakter obu funkcji wynika z oszacowanego parametru rozkładu Weibulla $\alpha_0 > 1$, co oznacza, że hazard zwiększa się wraz z upływem czasu. Wykres dobrze ilustruje zarówno zgodność modelu z danymi, jak i silny wpływ sprawności ECOG na ryzyko zgonu.

1.8 Zadanie 8

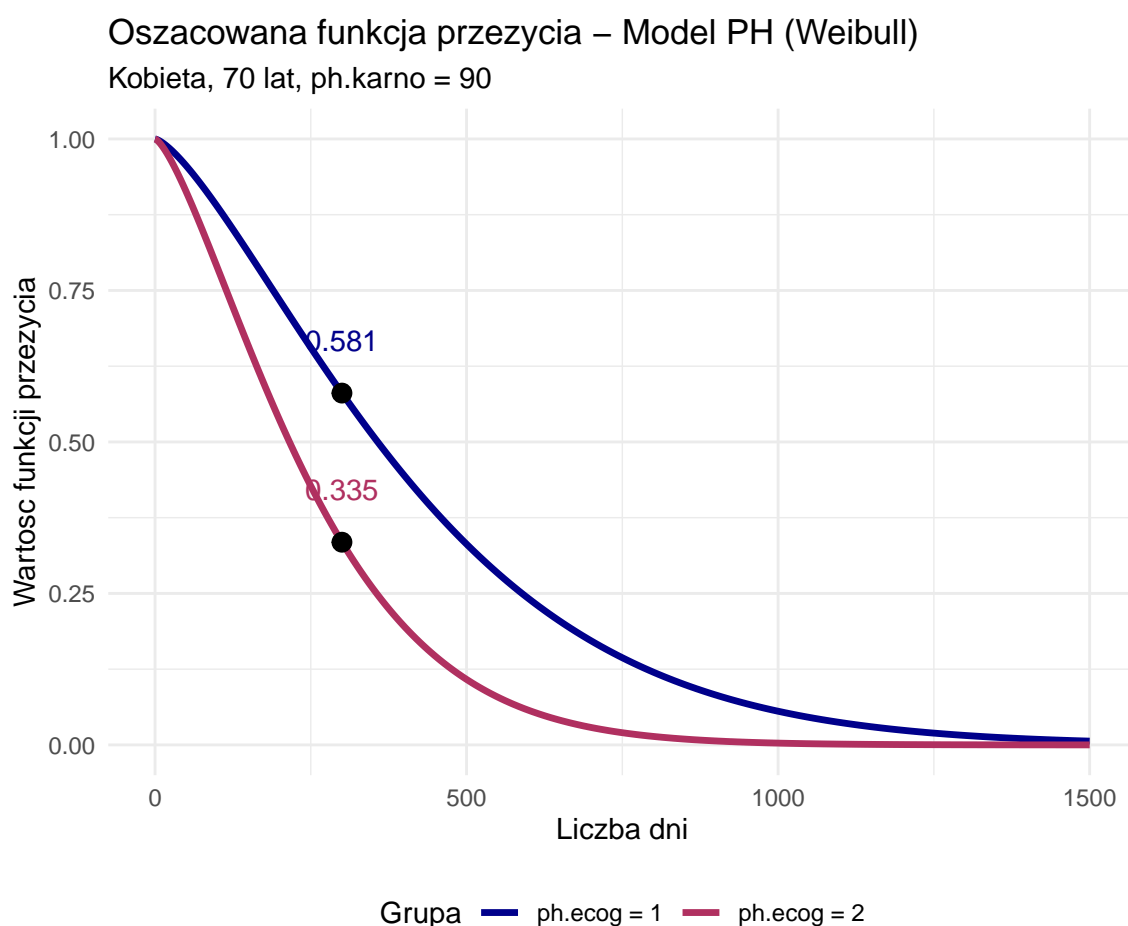
Kod

Poniższy kod definiuje funkcję S_PH, służącą do wyznaczania wartości funkcji przeżycia w modelu proporcjonalnych hazardów (PH) z rozkładem Weibulla.

```
S_PH <- function(x, z_vec) {  
  (exp(-lambda_PH * x^alpha_PH))^(exp(sum(beta_PH * z_vec)))  
}
```

Tabela 3: Szacowane prawdopodobieństwo przeżycia powyżej 300 dni (Model PH)

ph.ecog=1	0.5806085
ph.ecog=2	0.3345465



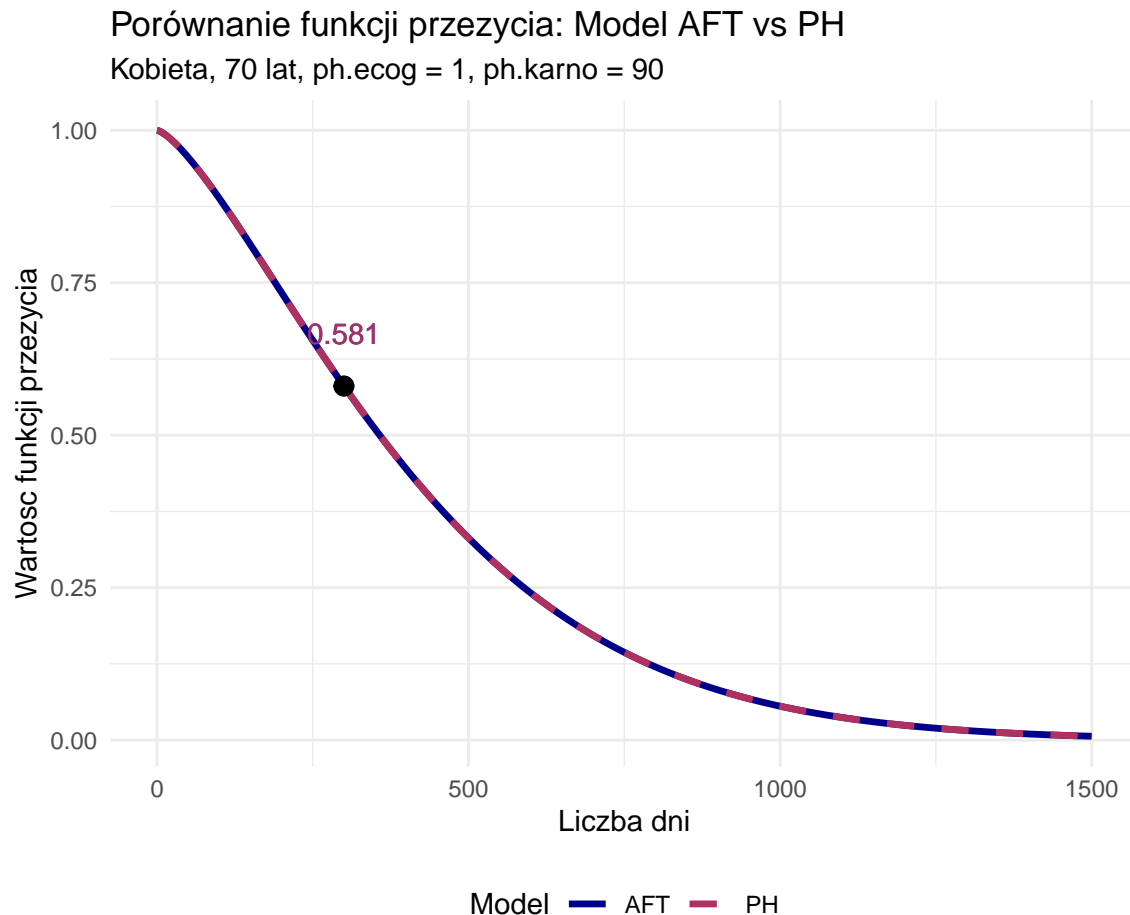
Rysunek 4: Oszacowana funkcja przeżycia w modelu PH (Weibull) dla kobiety w wieku 70 lat o ph.karno = 90, porównanie dla ph.ecog = 1 oraz ph.ecog = 2.

Tabela 3 przedstawia oszacowane prawdopodobieństwa przeżycia powyżej 300 dni dla dwóch grup pacjentek różniących się poziomem sprawności według skali ECOG, wyznaczone na pod-

stawie parametrycznego modelu proporcjonalnych hazardów Weibulla. Wyniki jednoznacznie wskazują, że kobieta w wieku 70 lat i $ph.karno = 90$ ma znacznie większą szansę przeżycia 300 dni, jeśli jej sprawność oceniana jest jako $ph.ecog = 1$ (prawdopodobieństwo przeżycia około 0.581), niż w przypadku $ph.ecog = 2$ (około 0.335). Oznacza to, że pogorszenie stanu funkcjonalnego przekłada się na wyraźny spadek przewidywanego czasu przeżycia.

Rysunek ?? ilustruje przebieg oszacowanych funkcji przeżycia dla tych samych dwóch grup. Krzywa odpowiadająca $ph.ecog = 1$ przebiega wyżej na całej długości osi czasu, co oznacza większe prawdopodobieństwo przeżycia w każdym momencie obserwacji. Z kolei krzywa dla $ph.ecog = 2$ opada szybciej, odzwierciedlając wyższy hazard i krótszy oczekiwany czas życia. Zaznaczone na wykresie punkty w $t = 300$ dodatkowo podkreślają różnicę między grupami, odpowiadając wartościom przedstawionym w Tabeli 3. Wspólnie tabela i wykres pokazują spójny obraz: gorsza sprawność ECOG istotnie obniża szanse długoterminowego przeżycia.

1.9 Zadanie 9



Rysunek 5: Porównanie oszacowanych funkcji przeżycia w modelach AFT i PH dla kobiety w wieku 70 lat o $ph.ecog = 1$ oraz $ph.karno = 90$.

Na Rysunku 5 przedstawiono porównanie oszacowanych funkcji przeżycia uzyskanych z dwóch modeli parametrycznych: modelu AFT (Weibulla) oraz modelu PH (Weibulla), dla kobiety w wieku 70 lat o $ph.ecog = 1$ i $ph.karno = 90$. Obie krzywe przebiegają niemal identycznie, co wskazuje, że w tym przypadku oba modele prowadzą do bardzo zbliżonych wniosków dotyczących przewidywanego czasu przeżycia. Zarówno model AFT, jak i model PH sugerują podobny kształt funkcji przeżycia oraz podobne wartości prawdopodobieństwa przeżycia w kolejnych momentach czasu.

Zaznaczony punkt w $t = 300$ dni pokazuje oszacowane prawdopodobieństwo przeżycia dla tej konkretnej pacjentki, które w obu modelach przyjmuje praktycznie tę samą wartość. Wykres dobrze ilustruje, że mimo odmiennych założeń teoretycznych modele AFT i PH mogą dawać bardzo zbliżone wyniki, jeśli struktura danych i zależności między zmiennymi są do siebie dopasowane.

2 Lista 10

2.1 Zadanie 1

W tej części analizy odchodzimy od założeń parametrycznych dotyczących rozkładu czasu życia na rzecz modelu semiparametrycznego - modelu proporcjonalnych hazardów Coxa. Model ten jest zdefiniowany poprzez funkcję hazardu o postaci:

$$h_z(t) = h_0(t) \exp(\beta^T z),$$

gdzie $h_0(t)$ jest bazową funkcją hazardu, która zależy od czasu, ale nie zależy od charakterystyk z , i o której nie przyjmujemy żadnych dodatkowych założeń dotyczących jej postaci. Część parametryczna modelu, $\exp(\beta^T z)$, zależy wyłącznie od wektora charakterystyk pacjenta i opisuje relatywny wpływ zmiennych objaśniających na ryzyko wystąpienia zdarzenia.

Odpowiadająca funkcja przeżycia ma postać:

$$S(t|z) = [S_0(t)]^{\exp(\beta^T z)},$$

gdzie $S_0(t)$ jest bazową funkcją przeżycia odpowiadającą $h_0(t)$, której postaci nie specyfikujemy w modelu.

Kluczową własnością modelu jest proporcjonalność hazardów, co oznacza, że stosunek hazardów dwóch jednostek o charakterystykach z_1 i z_2 nie zależy od czasu:

$$\frac{h(t|z_1)}{h(t|z_2)} = \exp(\beta^T(z_1 - z_2)).$$

Do modelu włączono następujące zmienne objaśniające:

- *age*: wiek w latach,
- *sex*: płeć (1-mężczyzna, 2-kobieta),
- *ph.ecog*: skala sprawności ECOG wg. lekarza (0-sprawność prawidłowa, 5-zgon),
- *ph.karno*: skala sprawności Karnofsky'ego wg. lekarza (100-sprawność prawidłowa, 0-zgon).

Współczynniki β w modelu Coxa interpretowane są poprzez ilorazy hazardów. Wartość $\exp(\beta_k)$ oznacza, ile razy zmienia się hazard zgonu przy wzroście k -tej zmiennej o jednostkę (przy pozostałych cechach stałych). Wartości większe od 1 oznaczają wzrost ryzyka zgonu, natomiast mniejsze od 1 jego spadek.

Kod

Do dopasowania modelu wykorzystano funkcję *coxph* z biblioteki *survival*.

```
model_cox <- coxph(Surv(time, status == 2) ~ age_cent + as.factor(sex) +
  as.factor(ph.ecog) + ph_karno_cent,
  data = dane)

beta <- model_cox$coeff
```

Tabela 4: Oszacowane parametry modelu proporcjonalnych hazardów Coxa

Zmienna	β
age_cent	0.0126
as.factor(sex)2	-0.5657
as.factor(ph.ecog)1	0.5781
as.factor(ph.ecog)2	1.2399
as.factor(ph.ecog)3	2.3959
ph_karno_cent	0.0124

2.2 Zadanie 2

Tabela 4 przedstawia oszacowane współczynniki modelu proporcjonalnych hazardów Coxa dla zmiennych opisujących pacjentów z danych lung. Każdy współczynnik β określa wpływ danej cechy na poziom hazardu, przy założeniu, że pozostałe zmienne pozostają stałe. Wartości dodatnie oznaczają wzrost hazardu, czyli skrócenie oczekiwanego czasu przeżycia, natomiast wartości ujemne jego spadek.

Analiza zmiennych wskazuje, że płeć żeńska wiąże się z ujemnym współczynnikiem β co oznacza, że przy stałych pozostałych charakterystykach, hazard zgonu dla kobiet jest niższy niż dla mężczyzn, co sprzyja dłuższemu przeżyciu. Najsilniejszy wpływ na model ma skala sprawności ECOG. Dodatnie parametry rosną gwałtownie wraz z pogorszeniem stanu pacjenta, co świadczy o drastycznym wzroście ryzyka u pacjentów o niższej sprawności. Wiek oraz skala Karnofsky’ego wykazują niewielkie dodatnie wartości β , co sugeruje marginalny wzrost hazardu wraz z każdym rokiem życia oraz przy wyższej punktacji Karnofsky’ego.

2.3 Zadanie 3

Kod

Najpierw obliczana jest bazowa skumulowana funkcja hazardu za pomocą funkcji *basehaz*, co daje estymator Breslowa dla $H_0(t)$. Następnie, korzystając z zależności $S_0(t) = \exp(-H_0(t))$, wyznaczana jest bazowa funkcja przeżycia.

```
H0_oszcz <- basehaz(model_cox, centered = TRUE)

S0_wartosci <- exp(-H0_oszcz$hazard)

S0_oszcz <- data.frame(
```

```

Czas = H0_oszcz$time,
S0 = S0_wartosci
)

wyniki_3 <- data.frame(
  Czas = H0_oszcz$time,
  H0_Breslow = H0_oszcz$hazard,
  S0_Breslow = S0_wartosci
)

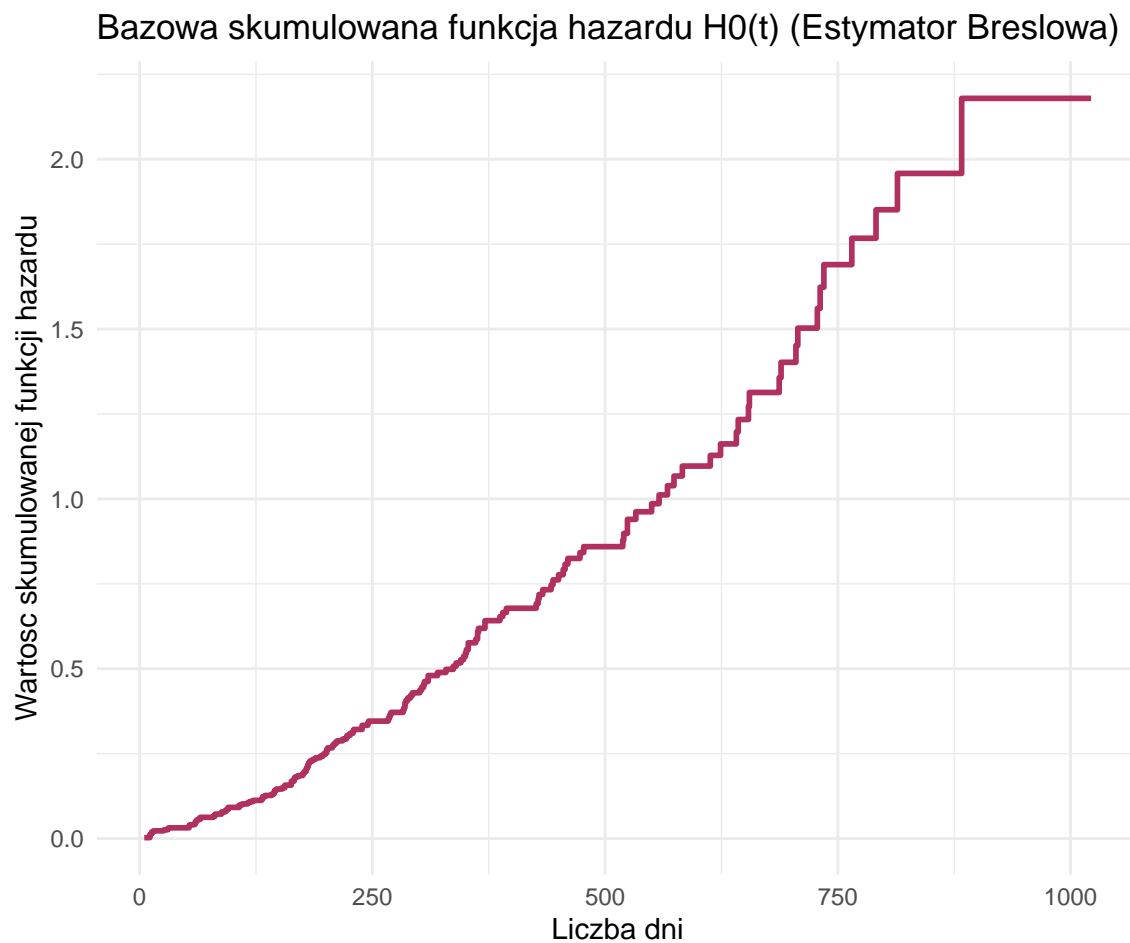
```

W Tabeli 5 oraz na Rysunkach 6 i 7 przedstawiono wyniki estymacji bazowych funkcji modelu Coxa, uzyskane przy pomocy estymatora Breslowa. Zastosowanie tej metody pozwala na wyznaczenie charakterystyk przeżycia bez konieczności przyjmowania konkretnego rozkładu bazowego, co jest kluczową cechą modeli semiparametrycznych.

Tabela 5: Oszacowanie bazowej skumulowanej funkcji hazardu i funkcji przeżycia (początkowe wartości)

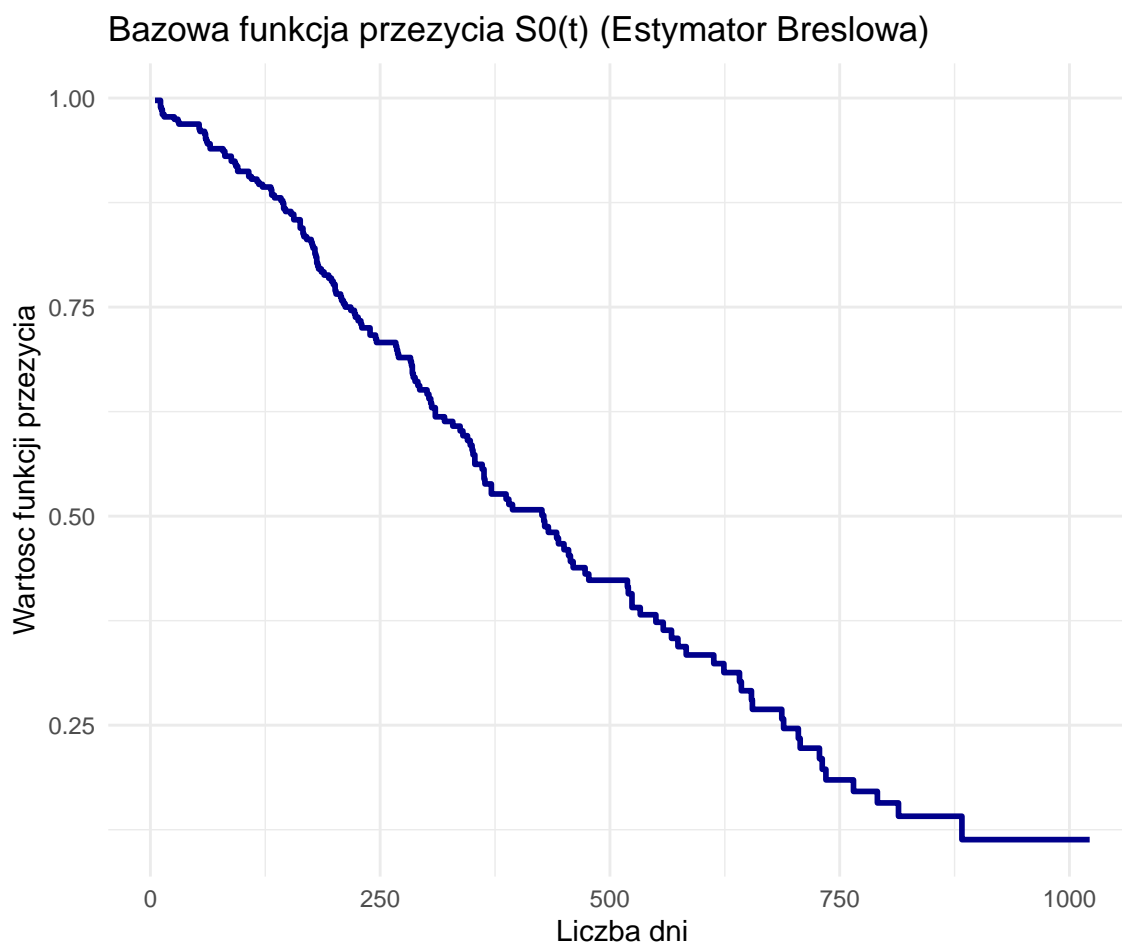
Czas	H0_Breslow	S0_Breslow
5	0.00279	0.99721
11	0.01123	0.98883
12	0.01408	0.98602
13	0.01986	0.98034
15	0.02278	0.97748
26	0.02571	0.97461
30	0.02867	0.97173
31	0.03167	0.96883
53	0.03770	0.96300
54	0.04075	0.96007

Tabela 5 zawiera pierwsze, początkowe wartości estymatora Breslowa skumulowanej funkcji hazardu $H_0(t)$ oraz odpowiadającej jej bazowej funkcji przeżycia $S_0(t)$. Już te wczesne punkty pokazują, że wartości $H_0(t)$ rosną powoli, a odpowiadające im wartości $S_0(t)$ pozostają bliskie 1, co odzwierciedla wysokie prawdopodobieństwo przeżycia na początku obserwacji.



Rysunek 6: Oszacowana bazowa skumulowana funkcja hazardu w modelu Coxa wyznaczona na podstawie estymatora Breslowa dla danych *lung*.

Rysunek 6 graficznie przedstawia przebieg skumulowanej funkcji hazardu. Jej charakterystyczny schodkowy kształt wynika z konstrukcji estymatora Breslowa, skoki pojawiają się wyłącznie w momentach, w których w danych odnotowano zdarzenia. Krzywa rośnie w sposób umiarkowany i regularny, co wskazuje, że ryzyko zgonu w populacji bazowej narasta stopniowo, bez gwałtownych zmian.



Rysunek 7: Oszacowana bazowa funkcja przeżycia w modelu Coxa, wyznaczona na podstawie estymatora Breslowa dla danych *lung*.

Konsekwencją rosnącego hazardu jest systematyczny spadek bazowej funkcji przeżycia $S_0(t)$, widoczny na Rysunku 7. Krzywa zaczyna się blisko wartości 1 i opada w miarę upływu czasu, zachowując płynny, monotoniczny przebieg. Jej kształt potwierdza, że w analizowanej populacji nie występują nagłe skoki ryzyka, a spadek przeżycia jest stopniowy i zgodny z typowym profilem danych klinicznych.

2.4 Zadanie 4

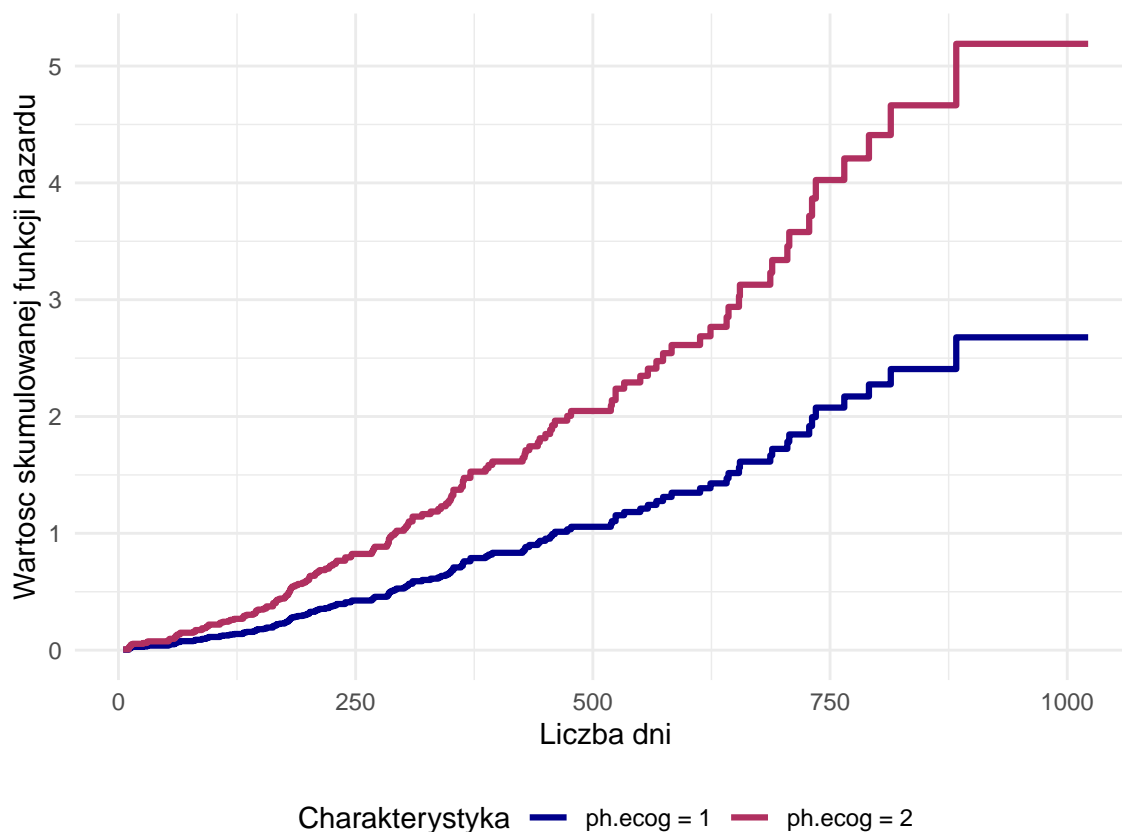
Kod

Kod oblicza skumulowany hazard dla dwóch profili pacjentów, mnożąc bazowy hazard przez czynnik $\exp(\beta^T z)$ odpowiedni dla każdego zestawu cech, a następnie zapisuje wektor czasów, w których te wartości zostały wyznaczone.

```
h_a <- H0_oszcz$hazard * exp(sum(beta * z_a))
h_b <- H0_oszcz$hazard * exp(sum(beta * z_b))
t <- H0_oszcz$time
```

Skumulowana funkcja hazardu – Model Coxa

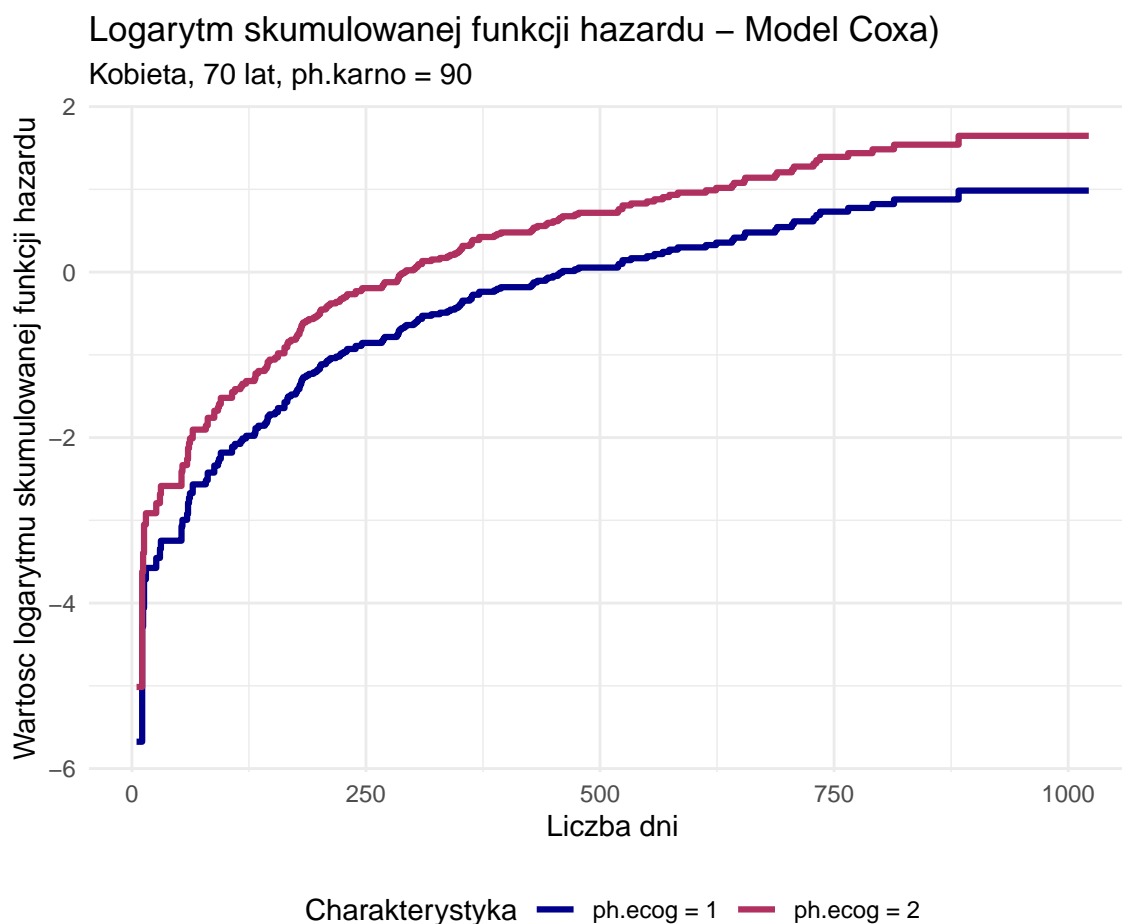
Kobieta, 70 lat, ph.karno = 90



Rysunek 8: Skumulowana funkcja hazardu oszacowana w modelu Coxa dla kobiety w wieku 70 lat o $ph.karno = 90$, porównanie dla $ph.ecog = 1$ oraz $ph.ecog = 2$.

Wykres przedstawiony na Rysunku 8 ilustruje przebieg skumulowanej funkcji hazardu w modelu Coxa dla dwóch poziomów sprawności według skali ECOG, przy założeniu, że analizowana osoba to kobieta w wieku 70 lat i o wartości $ph.karno$ równej 90. Obie krzywe mają charakter schodkowy, co wynika z konstrukcji estymatora Breslowa, skoki pojawiają się wyłącznie w momentach, w których w danych odnotowano zdarzenia.

Krzywa odpowiadająca poziomowi $ph.ecog = 2$ przebiega wyraźnie wyżej niż krzywa dla $ph.ecog = 1$, co oznacza, że skumulowany hazard rośnie szybciej dla pacjentów o gorszej sprawności. Innymi słowy, osoby z wyższą wartością ECOG mają większe narastające ryzyko zgonu w każdym momencie obserwacji.



Rysunek 9: Logarytm skumulowanej funkcji hazardu oszacowanej w modelu Coxa dla kobiety w wieku 70 lat o $ph.karno = 90$, porównanie dla $ph.ecog = 1$ oraz $ph.ecog = 2$.

Na Rysunku 9 przedstawiono wykresy logarytmów oszacowanych skumulowanych funkcji hazardu dla dwóch profili pacjentek, co stanowi kluczowe narzędzie weryfikacji założeń modelu proporcjonalnych hazardów Coxa.

Analiza wizualna tych krzywych pozwala na ocenę słuszności przyjętego modelu semiparametrycznego dla danych *lung*. Zgodnie z teorią modelu Coxa, jeśli założenie o proporcjonalności hazardów jest spełnione, krzywe odpowiadające różnym grupom powinny być względem siebie równoległe, a pionowa odległość między nimi powinna pozostawać stała w czasie i odpowiadać różnicy w wartościach liniowego predyktora $\beta^T z$.

Obserwując Rysunek 9, można stwierdzić, że obie linie, dla pacjentki o sprawności $ph.ecog = 1$ oraz $ph.ecog = 2$, zachowują niemal identyczny kształt na całej długości osi czasu. Stały odstęp między wykresami sugeruje, że iloraz hazardu dla tych dwóch grup nie ulega istotnym zmianom wraz z upływem dni obserwacji. Taki układ krzywych na wykresie logarytmicznym potwierdza zasadność zastosowania modelu Coxa i wskazuje, że nie ma wyraźnych przesłanek do podważenia założenia o proporcjonalności hazardów dla rozważanych charakterystyk.

Podsumowując, Rysunek 8 i 9 potwierdzają poprawność modelu Coxa. Na wykresie skumulo-

wanego hazardu (Rys. 8) widać wyraźną separację grup, a wykres logarytmiczny (Rys. 9) wykazuje niemal idealną równoległość krzywych. Ponieważ odstęp między liniami na skali logarytmicznej pozostaje stały w czasie, oznacza to, że iloraz hazardu jest niezmienny. Na podstawie tych wykresów nie ma powodów, by podważać poprawność modelu proporcjonalnych hazardów Coxa.

2.5 Zadanie 5

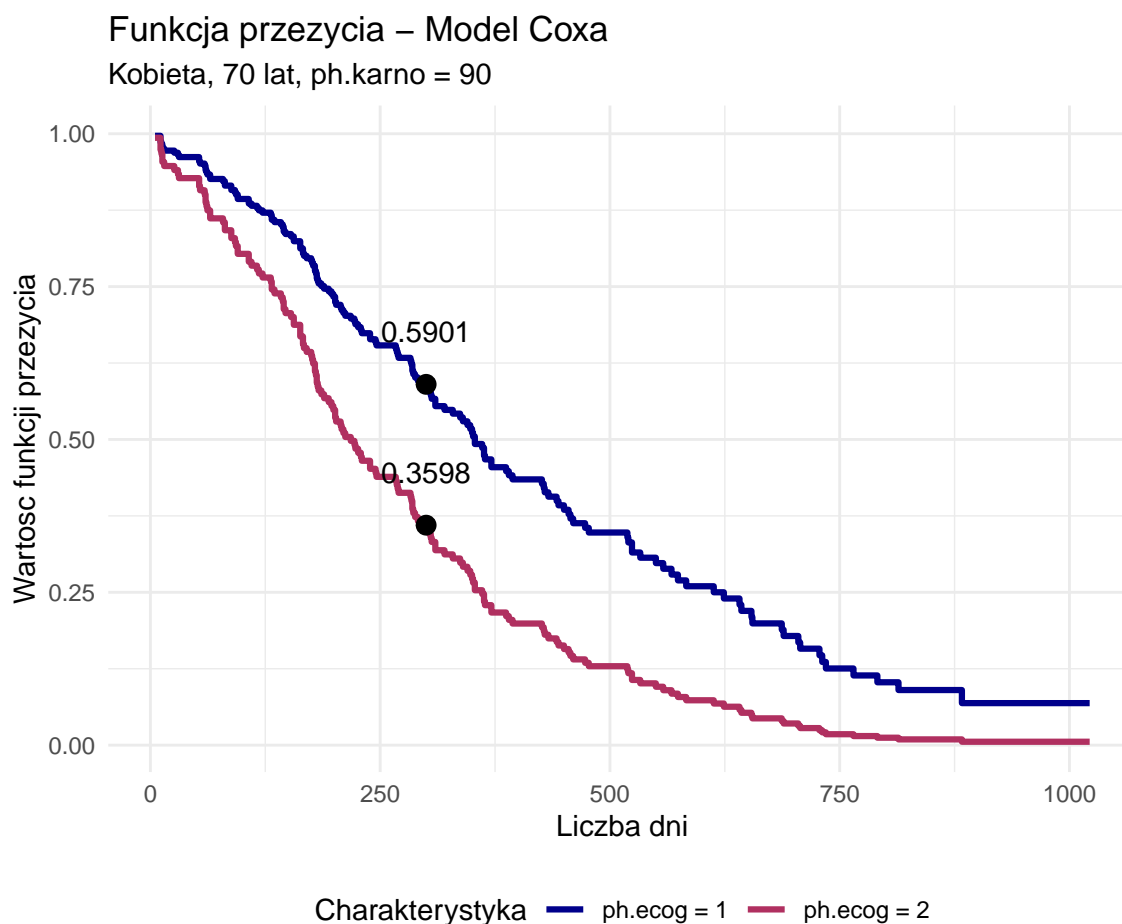
Kod

Wykorzystując zależność $S(t) = \exp(-H(t))$, kod oblicza funkcję przeżycia dla dwóch profili pacjentów używając wcześniej wyznaczonych wartości skumulowanego hazardu.

```
s_a <- exp(- h_a)
s_b <- exp(- h_b)
```

Tabela 6: Szacowane prawdopodobieństwo przeżycia powyżej 300 dni (Model Coxa)

Charakterystyka	Prawdopodobieństwo
ph.ecog = 1	0.5901344
ph.ecog = 2	0.3597682



Rysunek 10: Oszacowana funkcja przeżycia w modelu Coxa dla kobiety w wieku 70 lat o $ph.karno = 90$, porównanie dla $ph.ecog = 1$ oraz $ph.ecog = 2$, z zaznaczonymi wartościami w $t=300$ dni.

W Tabeli 6 oraz na Rysunku 10 przedstawiono oszacowane prawdopodobieństwa przeżycia dla dwóch profili pacjentek, wyznaczone w modelu Coxa przy użyciu estymatora Breslowa. Analiza dotyczy kobiet w wieku 70 lat o stałej wartości $ph.karno = 90$, różniących się stopniem sprawności fizycznej $ph.ecog$.

Zgodnie z danymi zawartymi w Tabeli 6, szacowana szansa na przeżycie powyżej 300 dni dla pacjentki o dobrej sprawności ($ph.ecog = 1$) wynosi około 0.5901. W przypadku pogorszenia kondycji fizycznej do stopnia $ph.ecog = 2$, prawdopodobieństwo to istotnie spada, osiągając wartość 0.3598.

Wizualizacja tych wyników na Rysunku 10 obrazuje przebieg funkcji przeżycia w czasie. Schodkowy kształt krzywych jest charakterystyczny dla nieparametrycznego estymatora bazowej funkcji przeżycia w modelu Coxa. Wykres potwierdza wyraźną dysproporcję między obiema grupami, krzywa dla pacjentki sprawniejszej przebiega konsekwentnie powyżej krzywej dla pacjentki o wyższym stopniu w skali ECOG. Zaznaczone na wykresie punkty w 300. dniu obserwacji odpowiadają wartościom numerycznym z Tabeli 6, podkreślając negatywny wpływ obniżonej sprawności na rokowanie pacjenta.

Tabela 7: Porównanie oszacowanych prawdopodobieństw przeżycia w $t = 300$ dniach dla modeli Cox, PH i AFT.

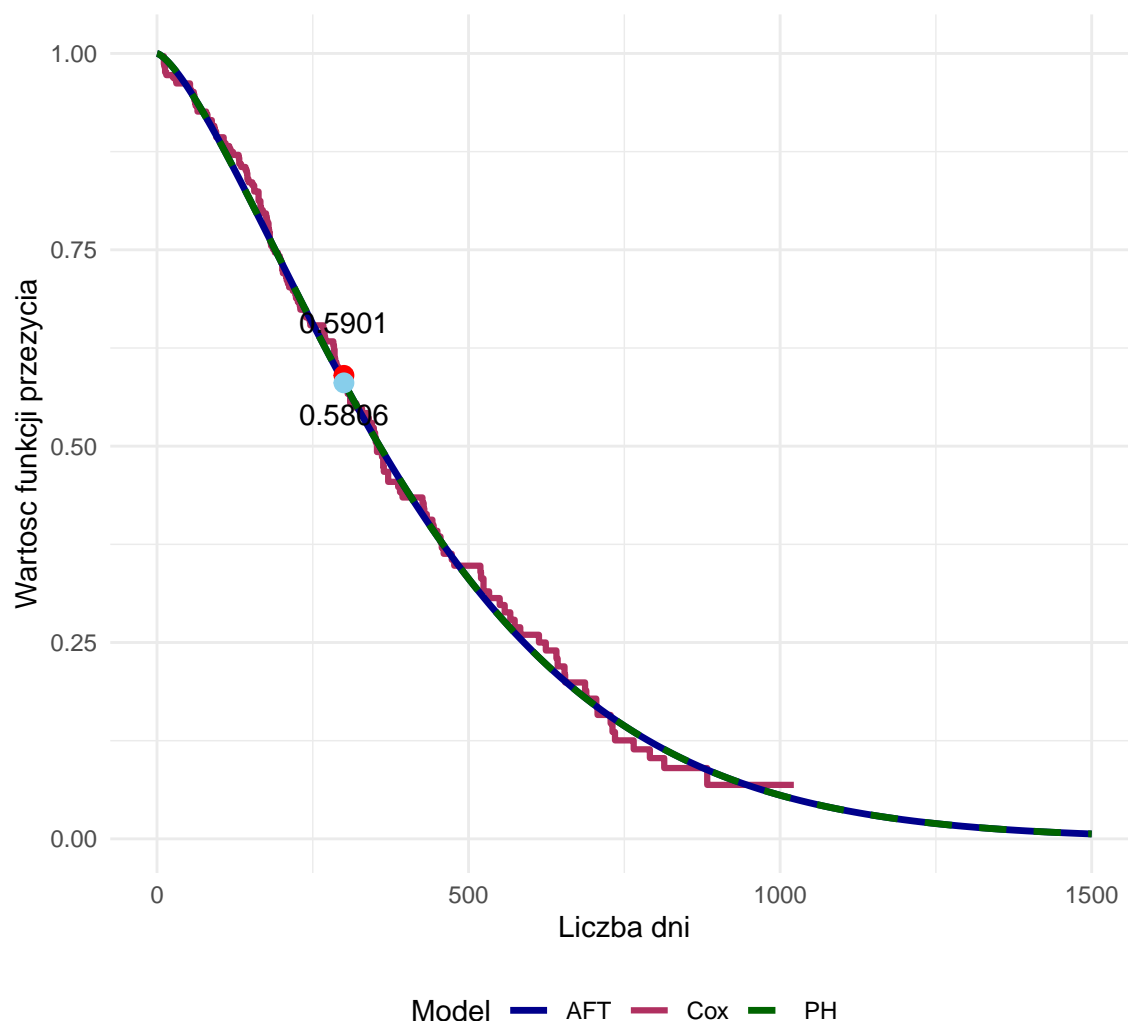
Model	Prawdopodobieństwo w $t = 300$
Cox (Breslow)	0.5901344
PH (Weibull)	0.5806085
AFT (Weibull)	0.5806085

Wyniki przedstawione w Tabeli 7 dla modeli parametrycznych PH oraz AFT opartych na rozkładzie Weibulla są identyczne i wynoszą 0,5806, natomiast otrzymane w modelu Coxa prawdopodobieństwo na poziomie 0,5901 jest bardzo zbliżone do nich. Nieznacznie wyższa wartość w modelu semiparametrycznym wynika z zastosowania estymatora Breslowa, który nie narzuca gładkiego kształtu funkcji, lecz elastycznie dopasowuje się do momentów wystąpienia zdarzeń w próbie. Tak wysoka zbieżność wszystkich trzech modeli potwierdza rzetelność przeprowadzonej analizy i dowodzi, że wybór metody modelowania ma w tym przypadku marginalny wpływ na finalne wnioski dotyczące rokowań pacjentki.

2.6 Zadanie 6

Porównanie funkcji przeżycia: Cox vs AFT vs PH

Kobieta, 70 lat, $ph.ecog = 1$, $ph.karno = 90$



Rysunek 11: Porównanie oszacowanych funkcji przeżycia dla modeli Coxa, AFT oraz PH dla kobiety w wieku 70 lat o $ph.ecog = 1$ i $ph.karno = 90$, z zaznaczonymi wartościami w $t = 300$ dni.

Na Rysunku 11 przedstawiono zbiorcze zestawienie estymowanych funkcji przeżycia dla pacjentki o wybranym profilu (kobieta, 70 lat, $ph.ecog = 1$, $ph.karno = 90$), porównując podejście semiparametryczne modelu Coxa z podejściem parametrycznym opartym na rozkładzie Weibulla (modele AFT i PH). Analiza wizualna pozwala zauważyć niemal całkowite pokrycie się krzywych odpowiadających modelom AFT i PH. W przeciwieństwie do gładkich linii modeli parametrycznych, funkcja przeżycia wyznaczona metodą Coxa z wykorzystaniem estymatora Breslowa ma charakterystyczną strukturę schodkową. Wynika to z jej nieparametrycznego charakteru oraz z tego, że zmienia się jedynie w momentach wystąpienia zdarzeń w danych.

Porównanie w punkcie $t = 300$ dni pokazuje, że model Coxa wskazuje na nieco wyższą szansę

przeżycia, wynoszącą 0.5901, podczas gdy modele Weibulla dają wynik na poziomie 0.5806. Mimo tej drobnej różnicy numerycznej, wszystkie trzy krzywe przebiegają bardzo blisko siebie na całym badanym dystansie czasowym.

Ostatecznie Rysunek 11 wykazuje, że wszystkie trzy podejścia opisują dane w bardzo podobny sposób, a różnice między modelami są marginalne i nie wpływają istotnie na interpretację przeżycia w analizowanej sytuacji.

3 Lista 11

3.1 Zadanie 1

Celem zadania jest oszacowanie parametrów modelu proporcjonalnych szans, przyjmując za zmienną zależną zmienną *time*, a za charakterystyki zmienne: *age*, *sex*, *ph.ecog*, *ph.karno*.

Zatem, do opisu zależności pomiędzy czasem przeżycia a zmiennymi objaśniającymi stosujemy semiparametryczny model proporcjonalnych szans. Stanowi on alternatywę dla modelu proporcjonalnych hazardów Coxa w sytuacjach, gdy założenie proporcjonalności hazardów nie jest spełnione.

Kluczowym podejściem w tym modelu jest szansa. Przy założeniu, że sukcesem jest wystąpienie zdarzenia przed czasem t , to szansę jednostki o wektorze charakterystyk z w chwili t definiujemy jako iloraz prawdopodobieństwa wystąpienia zdarzenia do prawdopodobieństwa przeżycia:

$$\theta_z(t) = \frac{1 - S_z(t)}{S_z(t)} \quad (1)$$

gdzie $S_z(t)$ jest funkcją przeżycia jednostki o wektorze charakterystyk z .

Semiparametryczny model proporcjonalnych szans zakłada, że szansa jednostki o charakterystyce z wyraża się wzorem:

$$\theta_z(t) = \theta_0(t) \exp(\beta^T z) \quad (2)$$

gdzie

$$\theta_0(t) = \frac{1 - S_0(t)}{S_0(t)} \quad (3)$$

oraz $S_0(t)$ jest bazową funkcją przeżycia (funkcją przeżycia jednostki o zerowym wektorze charakterystyk), a β jest wektorem współczynników modelu (nieznanych).

3.1.1 Kod

Implementacji kodu została oparta na zaproponowanej w uwadze do zadania funkcji `prop.odds` (zakładająca postać modelu proporcjonalnych szans zgodnego z wykładem).

```
#implementacja modelu
model_PS <- prop.odds(
  Event(time, status == 2) ~ age_cent + as.factor(sex) +
  as.factor(ph.ecog) + ph.karno_cent,
  data = dane,
  profile = 1
)
```

```
#szacowany parametr
beta <- as.numeric(model_PS$gamma)
```

Tabela 8: Oszacowane parametry modelu proporcjonalnych szans

Zmienna	β	wartość $\exp(\beta)$
age_cent	0.0132	1.0133
as.factor(sex)2	-0.9536	0.3853
as.factor(ph.ecog)1	0.5480	1.7298
as.factor(ph.ecog)2	1.4479	4.2543
as.factor(ph.ecog)3	1.9260	6.8623
ph_karno_cent	-0.0037	0.9963

3.2 Zadanie 2

Celem zadania jest interpretacja współczynników z zadania 3.1, przedstawionych w tabeli 8. Jak wskazuje nazwa modelu - iloraz szans dwóch jednostek o różnych wektorach charakterystyk z_1 i z_2 jest stały w czasie i zależy wyłącznie od różnicy tych charakterystyk:

$$\frac{\theta_{z_1}(t)}{\theta_{z_2}(t)} = \exp(\beta^T(z_1 - z_2)) \quad (4)$$

dzięki temu interpretacja współczynników jest bezpośrednia: wartość $\exp(\beta_k)$ informuje, jak zmienia się szansa wystąpienia zdarzenia przy wzroście k -tej zmiennej o jedną jednostkę.

Zatem z danych widocznych w Tabeli 8 możemy wyciągnąć następujące wnioski:

Dla zmiennej wieku (*agecent*) współczynnik β wynosi 0.0132, co daje iloraz szans 1.0133. Oznacza to, że z każdym kolejnym rokiem życia pacjenta (powyżej średniej), szansa zgonu nieznacznie wzrasta, przy ustalonych pozostałych parametrach. Jest to wpływ niekorzystny, ale umiarkowany.

Dla zmiennej płci (*sex*) (zakodowana binarnie (1 – mężczyzna, 2 – kobieta)). Współczynnik dla kobiet (*as.factor(sex)2*) jest ujemny i wynosi -0.9535 , a iloraz szans to 0.3854. Oznacza to, że kobiety mają wyraźnie mniejszą szansę zgonu w porównaniu do mężczyzn o tych samych pozostałych charakterystykach. Płeć żeńska jest zatem silnym czynnikiem ochronnym.

Zmienna *ph.ecog* jest zmienną kategoriową, gdzie poziomem odniesienia jest 0 (pełna sprawność). Wyniki pokazują drastyczny wzrost ryzyka wraz z pogorszeniem sprawności. Wraz ze wzrostem stopnia w skali ECOG dostrzegamy bardzo silny wpływ negatywny - szansa zgonu jest dla ECOG = 1 1.73, dla ECOG = 2 4.25, a dla ECOG = 3 nawet 6.85 razy wyższa w porównaniu do grupy referencyjnej. Widać zatem wyraźny trend - im wyższy stopień w skali ECOG, tym rokowanie jest gorsze.

Współczynnik *phkarnocent* jest ujemny (-0.0037), a iloraz szans wynosi 0.9963. Skala Karnofsky’ego działa odwrotnie do ECOG (wyższe wartości oznaczają lepszy stan zdrowia - 0

oznacza zgon, 100 sprawność prawidłową). Wynik ten interpretujemy tak, że wzrost wyniku w skali Karnofsky’ego o 1 punkt wiąże się ze spadkiem szansy zgonu.

Zatem możemy stwierdzić, że największy wpływ na ryzyko zgonu ma stan sprawności pacjenta (ECOG), a płeć żeńska znacząco poprawia rokowania.

3.3 Zadanie 3

Celem zadania jest wyznaczenie oszacowania bazowej skumulowanej funkcji hazardu i bazowej funkcji przeżycia odpowiadającej rozkładowi czasu życia w przyjętym modelu proporcjonalnych szans.

W modelu proporcjonalnych szans, podobnie jak w semiparametrycznym modelu proporcjonalnych hazardów Coxa, nie zakłada się z góry postaci rozkładu czasu życia. Model ten ma charakter semiparametryczny. Mamy następujące zależności wiążące poszukiwane funkcje z bazową szansą $\theta_0(t)$:

Bazowa funkcja przeżycia $S_0(t)$ jest związana z bazową szansą wzorem:

$$S_0(t) = \frac{1}{1 + \theta_0(t)} \quad (5)$$

Skumulowana funkcja przeżycia $H_0(t)$ wynika bezpośrednio z własności funkcji przeżycia:

$$H_0(t) = -\ln(S_0(t)) \quad (6)$$

3.3.1 Kod

Kod oparty jest na modelu zaimplementowanym w zadaniu 3.1.

```
#oszacowania bazowej funkcji przeżycia oraz skumulowanej funkcji hazardu
time0 <- model_PS$cum[, 1]
theta0 <- model_PS$cum[, 2]
S0_hat <- 1 / (1 + theta0)
H0_hat <- -log(S0_hat)
```

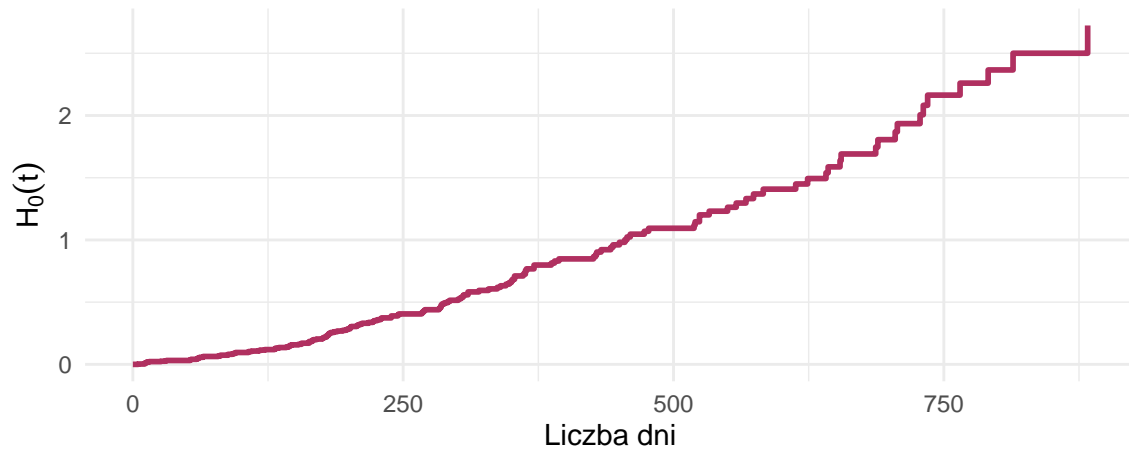
Tabela 9: 10 pierwszych oszacowań bazowej skumulowanej funkcji hazardu oraz bazowej funkcji przeżycia (model proporcjonalnych szans)

Czas	H0	S0
0.00000	0.00000	1.00000
5.00000	0.00266	0.99734
11.00000	0.00534	0.99467
11.00026	0.00807	0.99196
11.00036	0.01083	0.98923
12.00000	0.01361	0.98649

Czas	H0	S0
13.00000	0.01643	0.98370
13.00056	0.01932	0.98087
15.00000	0.02223	0.97802
26.00000	0.02516	0.97516

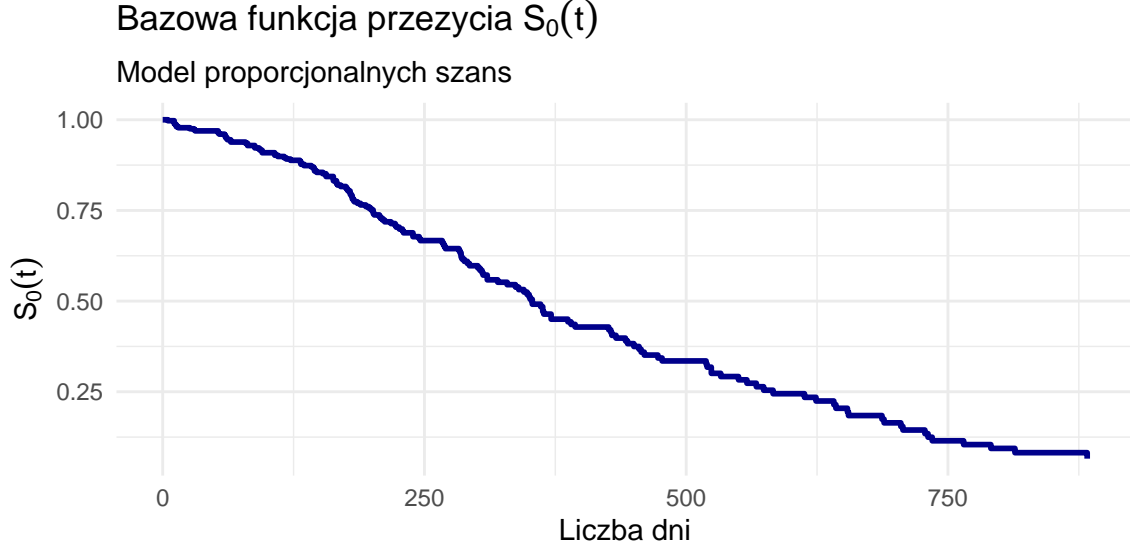
Bazowa skumulowana funkcja hazardu $H_0(t)$

Model proporcjonalnych szans



Rysunek 12: Oszacowana bazowa skumulowana funkcja hazardu w modelu proporcjonalnych szans.

Na Rysunku 12 przedstawiony jest wykres bazowej skumulowanej funkcji hazardu, która przyjmuje postać niemalejącej funkcji schodkowej. Wartość funkcji systematycznie rośnie, obrazując akumulację ryzyka wystąpienia zdarzenia (zgonu) w czasie dla pacjenta o referencyjnym poziomie cech.



Rysunek 13: Oszacowana bazowa funkcja przeżycia w modelu proporcjonalnych szans.

Na Rysunku 13 widoczny jest wykres bazowej funkcji przeżycia, której krzywa systematycznie maleje od wartości 1.0, odzwierciedlając dynamikę wymierania w grupie referencyjnej.

3.4 Zadanie 4

Celem zadania jest wyznaczenie oszacowania skumulowanej funkcji hazardu odpowiadającej rozkładowi czasu życia (w modelu proporcjonalnych szans) dla:

- (a) kobieta w wieku 70 lat, $ph.ecog = 1$, $ph.karno = 90$,
- (b) kobieta w wieku 70 lat, $ph.ecog = 2$, $ph.karno = 90$.

i porównania jej z wynikami z zadania 4 z listy 10.

W celu wyznaczenia skumulowanej funkcji hazardu dla konkretnego pacjenta w modelu proporcjonalnych szans, musimy skorzystać z relacji łączącej hazard z funkcją przeżycia. Funkcja przeżycia dla jednostki o wektorze charakterystyk z w tym modelu wyraża się wzorem:

$$S_z(t) = \frac{1}{1 + \theta_0(t) \exp(\beta^T z)} \quad (7)$$

gdzie $\theta_0(t)$ to bazowa szansa a β to oszacowane współczynniki modelu.

Skumulowana funkcja hazardu $H_z(t)$ jest zdefiniowana jako ujemny logarytm z funkcji przeżycia. Zatem dla modelu proporcjonalnych szans przyjmuje ona postać:

$$H_z(t) = -\ln \left(\frac{1}{1 + \theta_0(t) \exp(\beta^T z)} \right) = \ln \left(1 + \theta_0(t) \exp(\beta^T z) \right). \quad (8)$$

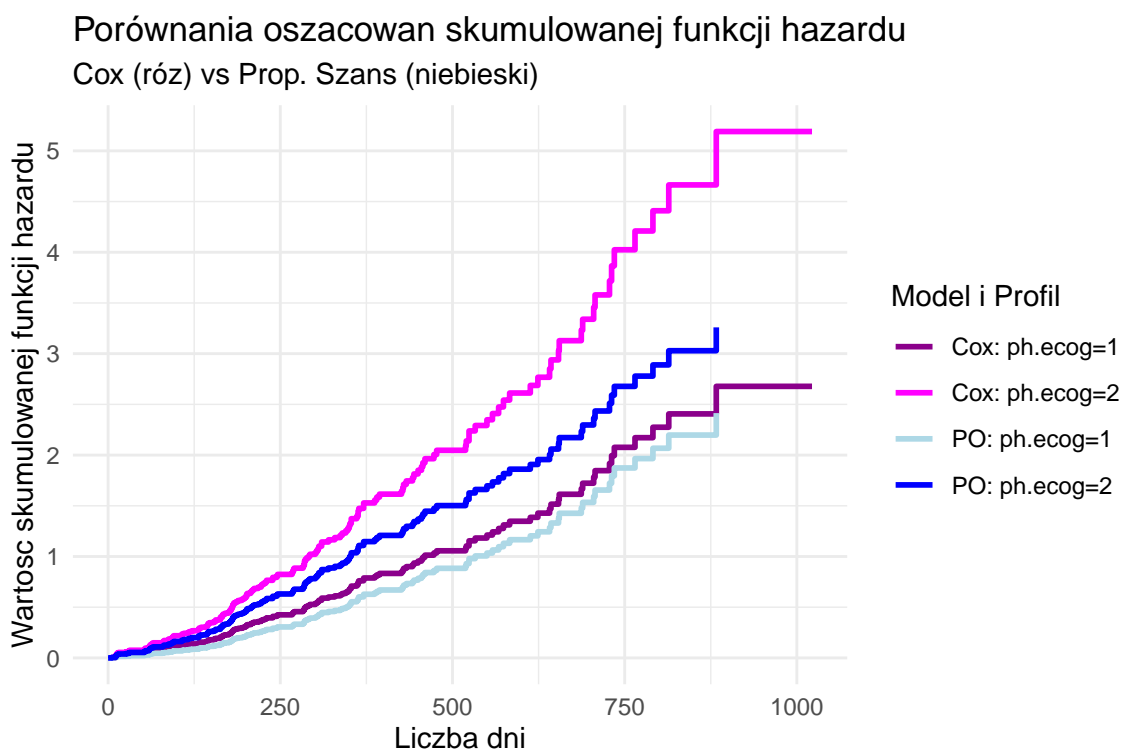
3.4.1 Kod

```
#zapisanie przypadku a oraz b (wektory charakterystyk)
z_a <- c(70 - age_m, 1, 1, 0, 0, 90 - ph_karno_m)
z_b <- c(70 - age_m, 1, 0, 1, 0, 90 - ph_karno_m)

#wyznaczenie funkcji przeżycia oraz skumulowanej funkcji hazardu
theta_0 <- (1 - wyniki_zad3$S0) / wyniki_zad3$S0

S_a_po <- 1 / (1 + theta_0 * exp(sum(beta * z_a)))
S_b_po <- 1 / (1 + theta_0 * exp(sum(beta * z_b)))

H_a_po <- -log(S_a_po)
H_b_po <- -log(S_b_po)
```



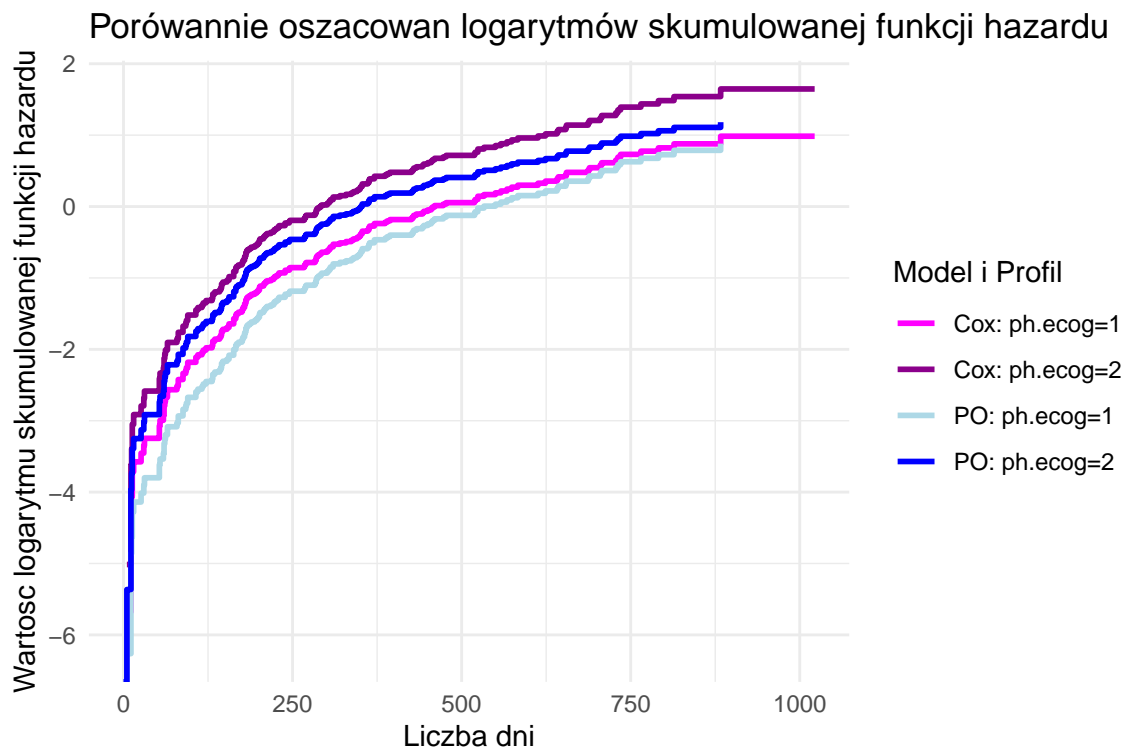
Rysunek 14: Wykres skumulowanej funkcji hazardu dla modelu proporcjonalnych szans oraz modelu Coxa (zadanie 4 z listy 10).

Na Rysunku 14 przedstawione są oszacowania skumulowanej funkcji hazardu dla modelu proporcjonalnych szans (odcienia koloru niebieskiego) oraz dla modelu Coxa (zrobione już wcześniej na liście 10 - zadanie 2.4) dla wcześniej wspomnianego przypadku a i b.

Na Rysunku 14 można zauważyć, że pacjentka o gorszym stanie sprawności ($ph.ecog = 2$) charakteryzuje się wyższymi wartościami skumulowanej funkcji hazardu w całym analizowanym przedziale czasu w porównaniu do pacjentki z $ph.ecog = 1$. Oznacza to większą skumulowaną

intensywność ryzyka zgonu oraz krótszy oczekiwany czas przeżycia dla pacjentek o wyższej wartości wskaźnika ECOG.

Porównując oba modele, można zauważyć, że model Coxa prowadzi do wyższych oszacowań skumulowanej funkcji hazardu niż model proporcjonalnych szans, szczególnie dla pacjentki o $ph.ecog = 2$. Różnice pomiędzy krzywymi są niewielkie w początkowym okresie obserwacji, natomiast narastają wraz z upływem czasu, co wskazuje na odmienny sposób modelowania ryzyka w obu podejściach.



Rysunek 15: Wykres logarytmu skumulowanej funkcji hazardu dla modelu proporcjonalnych szans oraz modelu Coxa (zadanie 4 z listy 10).

Na Rysunku 15 przedstawione są oszacowania logarytmu skumulowanej funkcji hazardu dla modelu proporcjonalnych szans (odcienie koloru niebieskiego) oraz dla modelu Coxa (zrobione już wcześniej na liście 10 - zadanie 2.4) dla wcześniej wspomnianego przypadku a i b.

Dla modeli widocznych na Rysunku 15 widoczna jest wyraźna separacja między krzywymi pacjentki A i B — pacjentka o gorszej sprawności ($ECOG = 2$) ma wyższy logarytm skumulowanego hazardu, co oznacza większe ryzyko zgonu.

W modelu Coxa krzywe są niemal równoległe, co potwierdza założenie o proporcjonalności hazardów — różnice między grupami są stałe w czasie. W modelu proporcjonalnych szans krzywe zbliżają się do siebie, co odzwierciedla malejące różnice w ryzyku zgonu wraz z upływem czasu.

3.5 Zadanie 5

Celem zadania jest wyznaczenie oszacowania funkcji przeżycia odpowiadającej rozkładowi czasu życia (w modelu proporcjonalnych szans) dla:

- (a) kobieta w wieku 70 lat, $ph.ecog = 1$, $ph.karno = 90$,
- (b) kobieta w wieku 70 lat, $ph.ecog = 2$, $ph.karno = 90$.

Na podstawie wyznaczonych funkcji należy obliczyć szacowane prawdopodobieństwo przeżycia powyżej 300 dni dla obu przypadków. Dodatkowo, uzyskany wynik dla profilu (a) zostanie porównany z prawdopodobieństwem obliczonym w analogicznym zadaniu dla modelu Coxa (Lista 10, Zadanie 5).

Znając część teoretyczną z zadania 3.4 możemy przejść do kodu i interpretacji wyników.

3.5.1 Kod

```
#czas dla którego szacujemy
t_cel <- 300
#znalezienie najbliższego indeksu czasu w wynikach dla modelu prop. szans
idx_300 <- which.min(abs(wyniki_zad3$Czas - t_cel))
rzeczywisty_czas <- wyniki_zad3$Czas[idx_300]
#prawdopodobieństwa dla obu profili
prawd_a_300 <- S_a_po[idx_300]
prawd_b_300 <- S_b_po[idx_300]
```

Tabela 10: Szacowane prawdopodobieństwo przeżycia powyżej 300 dni - porównanie modeli

Model	Profil pacjentki	Prawdopodobieństwo $P(T > 300)$
Prop. Szans (PO)	Kobieta (a): $ph.ecog = 1$	0.6692
Prop. Szans (PO)	Kobieta (b): $ph.ecog = 2$	0.4513
Cox (PH) [Lista 10]	Kobieta (a): $ph.ecog = 1$	0.5901

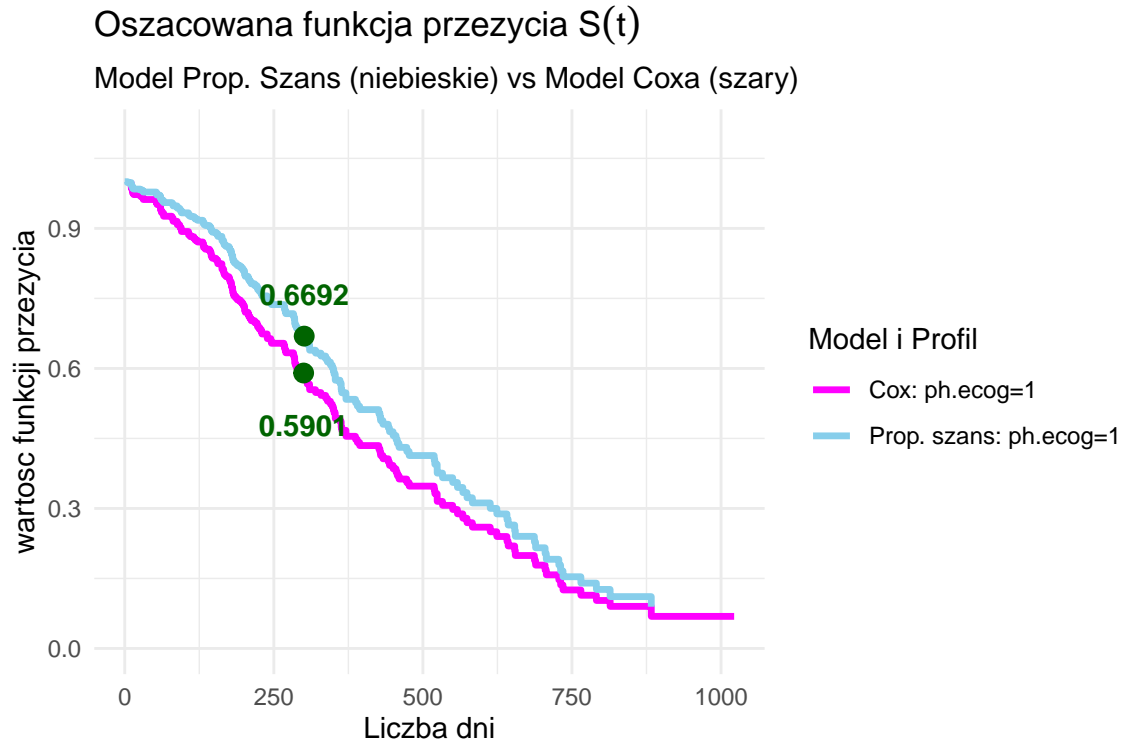
Tabela 10 zawiera prawdopodobieństwo przeżycia powyżej 300 dni dla dwóch profili pacjentek w modelu proporcjonalnych szans oraz pacjentki a w modelu Coxa.

W ramach modelu proporcjonalnych szans widoczna jest wyraźna różnica w rokowaniach zależna od stanu zdrowia. Pacjentka o lepszej sprawności ($ph.ecog = 1$) ma wyraźnie wyższe prawdopodobieństwo przeżycia niż pacjentka z gorszą sprawnością ($ph.ecog = 2$). Potwierdza to, że każdy stopień pogorszenia w skali ECOG jest silnym czynnikiem ryzyka.

Porównując wyniki uzyskane w modelu proporcjonalnych szans z rezultatami otrzymanymi w modelu Coxa (lista 10), można zauważyć, że dla pacjentki o $ph.ecog = 1$ model proporcjonalnych szans prowadzi do wyższego oszacowania prawdopodobieństwa przeżycia powyżej 300 dni (daje bardziej optymistyczne oszacowanie) niż model Coxa.

3.6 Zadanie 6

Celem zadania jest narysowanie wykresu i porównanie modeli Coxa (Lista 10 zadanie 2.5) z modelem proporcjonalnych szans dla pacjentki *a*.



Rysunek 16: Wykres oszacowanej funkcji przeżycia dla modelu proporcjonalnych szans oraz pacjentki *a* modelu Coxa (zadanie 4 z listy 10).

Rysunek 16 przedstawia oszacowania funkcji przeżycia dla pacjentki *a* z modelu proporcjonalnych szans oraz pacjentki *a* z modelu Coxa.

Krzywa modelu proporcjonalnych szans (jasnoniebieska) przebiega konsekwentnie powyżej krzywej modelu Coxa (różowa). Wskazuje to, że przy przyjętych założeniach model proporcjonalnych szans generuje w tym przypadku bardziej optymistyczne rokowanie (wolniejszy spadek funkcji przeżycia) niż model proporcjonalnych hazardów.

Wnioski te sugerują, że wybór modelu może wpływać na interpretację rokowania klinicznego.

4 Lista 12

4.1 Zadanie 1

Celem zadania jest weryfikacja hipotez statystycznych dotyczących istotności poszczególnych zmiennych objaśniających w parametrycznym modelu przyspieszonego czasu awarii (AFT) z założonym rozkładem Weibulla dla bazowej funkcji przeżycia. Analiza zostanie przeprowadzona na zbiorze danych lung. Weryfikacja obejmuje:

- zmienną *age* (test Walda oraz test IW na poziomie istotności 0.05),
- zmienną *sex* (test Walda oraz test IW na poziomie istotności 0.05),
- zmienną *ph.ecog* (tylko test Walda na poziomie istotności 0.05).

Teoria dotycząca modelu przyspieszonego czasu awarii (AFT) jest przedstawiona w zadaniu 1.1.

Skupmy się zatem na metodach testowania istotności parametrów modeli. W celu istotności zmiennych objaśniających w modelu między innymi AFT stosuje się klasyczne testy statystyczne oparte na funkcji wiarygodności - test Walda oraz test IW.

Ocena formalnie sprowadza się to do weryfikacji hipotezy zerowej postaci:

$$H_0 : \beta_k = 0 \quad (9)$$

przy hipotezie alternatywnej

$$H_1 : \beta_k \neq 0 \quad (10)$$

gdzie β_k jest współczynnikiem regresji przy k -tej zmiennej. Odrzucenie hipotezy zerowej oznacza, że dana zmienna jest istotna statystycznie w modelu.

Statystyka testowa Walda opiera się na ocenie odległości estymatora $\hat{\beta}$ od wartości hipotetycznej (zazwyczaj 0), ważonej przez błąd standardowy estymacji. Dla pojedynczego parametru statystyka Z wyraża się wzorem:

$$Z = \frac{\hat{\beta} - \beta_0}{\sqrt{Var(\hat{\beta})}} \quad (11)$$

gdzie przy założeniu prawdziwości H_0 , statystyka ta ma asymptotyczny rozkład normalny standardowy $N(0, 1)$. Wartość poziomu krytycznego (p -value) jest postaci:

$$p = 2(1 - \Phi(|z|)) \quad (12)$$

Test IW porównuje dopasowanie dwóch modeli: modelu pełnego (\mathcal{M}_2) oraz modelu zagnieżdżonego (\mathcal{M}_1), w którym pewne parametry (np. współczynnik przy badanej zmiennej) są równe zero. Statystyka testowa Λ_n zdefiniowana jest jako:

$$\Lambda_n(X) = -2 \ln \left(\frac{L(\mathcal{M}_1)}{L(\mathcal{M}_2)} \right) = -2[\ln L(\mathcal{M}_1) - \ln L(\mathcal{M}_2)] \quad (13)$$

gdzie $L(\cdot)$ to funkcja wiarygodności. Przy założeniu prawdziwości hipotezy zerowej, statystyka ta ma asymptotyczny rozkład χ^2 z liczbą stopni swobody równą różnicy liczby parametrów w obu modelach.

4.1.1 Kod

```
#model pełny
model_AFT <- survreg(Surv(time, status) ~ age_cent + as.factor(sex) +
                     as.factor(ph.ecog) + ph_karno_cent,
                     data = dane, dist = "weibull")

#podpunkt a)
#model zredukowany (bez age)
model_AFT_bez_age <- survreg(Surv(time, status) ~ as.factor(sex) +
                             as.factor(ph.ecog) + ph_karno_cent,
                             data = dane, dist = "weibull")

#test Walda
p_wald_AFT_age <- summary(model_AFT)$table["age_cent", "p"]

#test IW
test_iw_AFT_age <- anova(model_AFT_bez_age, model_AFT)
p_iw_AFT_age <- test_iw_AFT_age$`Pr(>Chi)`[2]

#podpunkt b)
#model zredukowany (bez sex)
model_AFT_bez_sex <- survreg(Surv(time, status) ~ age_cent +
                             as.factor(ph.ecog) + ph_karno_cent,
                             data = dane, dist = "weibull")

#test Walda
p_wald_AFT_sex <- summary(model_AFT)$table["as.factor(sex)2", "p"]

#test IW
test_iw_AFT_sex <- anova(model_AFT_bez_sex, model_AFT)
p_iw_AFT_sex <- test_iw_AFT_sex$`Pr(>Chi)`[2]

#podpunkt c)
#model zredukowany (bez ph.ecog)
model_AFT_bez_ecog <- survreg(Surv(time, status) ~
                              age_cent + as.factor(sex) +
```

```

      ph.karno.cent, data = dane,
      dist = "weibull")

#test IW
test_iw_AFT_ecog <- anova(model_AFT_bez_ecog, model_AFT)
p_iw_AFT_ecog <- test_iw_AFT_ecog$`Pr(>Chi)`[2]

```

Tabela 11: Wyniki testów istotności dla modelu AFT

Zmienna	P-value (Wald)	P-value (IW)
age	0.20530	0.20137
sex	0.00091	0.00060
ph.ecog (całość)	NA	0.00214

W Tabeli 11 widoczne są wyniki przeprowadzonych testów Walda oraz ilorazu wiarygodności (IW), przy przyjętym poziomie istotności $\alpha = 0.05$.

Dla zmiennej *age* zarówno test Walda ($p - value \approx 0.205$), jak i test IW ($p - value \approx 0.201$) wskazują na brak podstaw do odrzucenia hipotezy zerowej. Oznacza to, że wiek pacjenta nie jest istotnym predyktorem czasu przeżycia w analizowanym modelu.

Dla zmiennej *sex* w obu testach uzyskano wartości p znacznie poniżej poziomu istotności ($p - value < 0.001$), co jednoznacznie przemawia za odrzuceniem hipotezy zerowej. Płeć jest statystycznie istotnym czynnikiem wpływającym na czas trwania życia.

Dla zmiennej *ph.ecog* test ilorazu wiarygodności przeprowadzony dla zmiennej kategorycznej jako całości dał wynik $p - value \approx 0.002$. Pozwala to na odrzucenie hipotezy zerowej o równości rozkładów w grupach i potwierdza, że stopień sprawności ECOG różnicuje czas przeżycia pacjentów w sposób istotny statystycznie.

4.2 Zadanie 2

Celem zadania jest weryfikacja hipotez statystycznych dotyczących istotności poszczególnych zmiennych objaśniających w parametrycznym modelu proporcjonalnych hazardów Coxa (PH). Analiza zostanie przeprowadzona na zbiorze danych lung. Weryfikacja obejmuje:

- zmienną *age* (test Walda oraz test IW na poziomie istotności 0.05),
- zmienną *sex* (test Walda oraz test IW na poziomie istotności 0.05),
- zmienną *ph.ecog* (tylko test Walda na poziomie istotności 0.05).

Teoria dotycząca modelu proporcjonalnych hazardów Coxa (PH) jest przedstawiona w zadaniu 1.5. Teoria dotycząca metod testowania istotności parametrów modeli jest przedstawiona w zadaniu 4.1.

4.2.1 Kod

```
#model pełny
model_Cox <- coxph(Surv(time, status == 2) ~ age_cent +
  as.factor(sex) + as.factor(ph.ecog)
  + ph_karno_cent, data = dane)

#podpunkt a)
#model zredukowany (bez age)
model_Cox_bez_age <- coxph(Surv(time, status == 2) ~ as.factor(sex)
  + as.factor(ph.ecog)
  + ph_karno_cent, data = dane)

#test Walda
p_wald_Cox_age <- summary(model_Cox)$
  coefficients["age_cent", "Pr(>|z|)"]

#test IW
test_iw_Cox_age <- anova(model_Cox_bez_age, model_Cox, test = "LRT")
p_iw_Cox_age <- test_iw_Cox_age$`Pr(>|Chi|)`[2]

#podpunkt b)
#model zredukowany (bez sex)
model_Cox_bez_sex <- coxph(Surv(time, status == 2) ~ age_cent +
  as.factor(ph.ecog) +
  ph_karno_cent, data = dane)

#test Walda
p_wald_Cox_sex <- summary(model_Cox)$
  coefficients["as.factor(sex)2", "Pr(>|z|)"]

#test IW
test_iw_Cox_sex <- anova(model_Cox_bez_sex, model_Cox, test = "LRT")
p_iw_Cox_sex <- test_iw_Cox_sex$`Pr(>|Chi|)`[2]

#podpunkt c)
#model zredukowany (bez ph.ecog)
model_Cox_bez_ecog <- coxph(Surv(time, status == 2) ~
  age_cent + as.factor(sex) +
  ph_karno_cent, data = dane)
```

```
#test IW
test_iw_Cox_ecog <- anova(model_Cox_bez_ecog, model_Cox, test = "LRT")
p_iw_Cox_ecog <- test_iw_Cox_ecog$`Pr(>|Chi|)`[2]
```

Tabela 12: Wyniki testów istotności dla modelu proporcjonalnych hazardów Coxa

Zmienna	P-value (Wald)	P-value (IW)
age	0.18390	0.18038
sex	0.00086	0.00063
ph.ecog	NA	0.00365

W Tabeli 12 widoczna są wyniki analizy statystycznej przeprowadzonej dla modelu proporcjonalnych hazardów Coxa, przy poziomie istotności $\alpha = 0.05$,

Dla zmiennej *age* zarówno test Walda ($p - value \approx 0.184$), jak i test ilorazu wiarygodności ($p - value \approx 0.180$) wskazują na brak istotności statystycznej wieku. Podobnie jak w modelu AFT, nie ma podstaw do stwierdzenia, że wiek wpływa na ryzyko zgonu w analizowanej grupie.

Dla zmiennej *sex* uzyskane wartości $p - value$ są bardzo bliskie zeru ($p < 0.001$) w obu testach. Potwierdza to, że płeć jest silnym i istotnym predyktorem wpływającym na ryzyko zgonu.

Dla zmiennej *ph.ecog* test ilorazu wiarygodności dał wynik $p - value \approx 0.004$. Oznacza to odrzucenie hipotezy o braku wpływu stopnia sprawności na przeżycie. Różnice między poziomami tej zmiennej są istotne statystycznie.

4.3 Zadanie 3

Celem zadania jest przeprowadzenie procedury doboru zmiennych (selekcji modelu) dla parametrycznego modelu przyspieszonego czasu awarii (AFT) z rozkładem Weibulla. Spośród dostępnych charakterystyk (*age*, *sex*, *ph.ecog4*, *ph.karno*) należy wyłonić optymalny podzbiór zmiennych, które najlepiej opisują czas przeżycia pacjentów.

Zadanie wymaga porównania trzech różnych metod selekcji: Manualnej metody eliminacji wstecznej, opartej na wartościach $p - value$ z testu ilorazu wiarygodności (IW). Automatycznej selekcji krokowej w oparciu o kryterium informacyjne Akaike'a (AIC). Automatycznej selekcji krokowej w oparciu o Bayesowskie kryterium informacyjne (BIC).

Metoda eliminacji wstecznej jest to iteracyjna procedura oparta na testach istotności. *Model początkowy*: Model pełny, zawierający wszystkie rozważane zmienne. *Krok*: Weryfikujemy hipotezy o nieistotności poszczególnych zmiennych (korzystając z testu ilorazu wiarygodności). Usuwamy zmienną, dla której wartość poziomu krytycznego ($p - value$) jest największa i przekracza ustalony próg (przyjmujemy próg np. $\alpha = 0.15$). *Stop*: Procedurę kończymy, gdy nie można usunąć żadnej zmiennej (wszystkie pozostałe są istotne).

Kryteria Informacyjne (AIC i BIC) to metody oceniające jakość modelu poprzez funkcję wiarygodności L , nakładając “karę” za liczbę estymowanych parametrów k . Celem jest minimalizacja wartości kryterium.

$$AIC = -2 \ln L + 2k \quad (14)$$

Kara za złożoność wynosi $2k$. AIC preferuje modele, które dobrze przewidują przyszłe dane, ale ma tendencję do wybierania modeli nieco bardziej złożonych przy dużych próbach.

$$BIC = -2 \ln L + k \ln n \quad (15)$$

Kara za złożoność wynosi $k \ln n$. Ponieważ dla liczebności próby $n > 7$ zachodzi $\ln n > 2$, kara w BIC jest surowsza niż w AIC. Skutkuje to wybieraniem modeli prostszych (z mniejszą liczbą zmiennych).

4.3.1 Kod - podpunkt a

Kod do podpunktu a) zakłada zastosowanie metody eliminacji wstecznej. Zatem mamy dla kroku 1:

```
#model pełny
model.0 <- survreg(Surv(time, status) ~ age_cent + as.factor(sex)
                  + as.factor(ph.ecog) + ph_karno_cent,
                  data = dane, dist = "weibull")

#modele zredukowane (bez jednej zmiennej)
model.bez_age <- survreg(Surv(time, status) ~ as.factor(sex)
                       + as.factor(ph.ecog) + ph_karno_cent,
                       data=dane, dist="weibull")

model.bez_sex <- survreg(Surv(time, status) ~ age_cent
                       + as.factor(ph.ecog) + ph_karno_cent,
                       data=dane, dist="weibull")

model.bez_ecog <- survreg(Surv(time, status) ~ age_cent
                        + as.factor(sex) +
                        ph_karno_cent, data=dane, dist="weibull")

model.bez_karno <- survreg(Surv(time, status) ~ age_cent
                        + as.factor(sex) + as.factor(ph.ecog),
                        data=dane, dist="weibull")

#p-value dla IW
p_value_k1 <- c(
  anova(model.0, model.bez_age)[2, "Pr(>Chi)"],
```



```
anova(model.0, model.bez_sex)[2, "Pr(>Chi)"],
anova(model.0, model.bez_ecog)[2, "Pr(>Chi)"],
anova(model.0, model.bez_karno)[2, "Pr(>Chi)"])
```

Tabela 13: Krok 1: Weryfikacja zmiennych w modelu pełnym

Zmienna usuwana	P-value (IW)
age	0.2014
sex	0.0006
ph.ecog	0.0021
ph.karno	0.1332

Jak widać w Tabeli 13 zmienna *age* ma największą wartość poziomu krytycznego ($p\text{-value} \approx 0.2014$). Ponieważ wartość ta przekracza przyjęty próg istotności $\alpha = 0.15$, zmienną *age* możemy uznać za nieistotną i usunąć ją z modelu. Otrzymujemy, więc następujący krok 2:

```
#model po usunięciu age
model.0_k2 <- survreg(Surv(time, status) ~ as.factor(sex)
+ as.factor(ph.ecog) + ph.karno_cent,
data = dane, dist = "weibull")

#modele zredukowane
model.bez_sex_k2 <- survreg(Surv(time, status) ~
as.factor(ph.ecog) + ph.karno_cent,
data=dane, dist="weibull")

model.bez_ecog_k2 <- survreg(Surv(time, status) ~
as.factor(sex) + ph.karno_cent,
data=dane, dist="weibull")

model.bez_karno_k2 <- survreg(Surv(time, status) ~
as.factor(sex) + as.factor(ph.ecog),
data=dane, dist="weibull")

#wartości p-value
p_value_k2 <- c(
anova(model.0_k2, model.bez_sex_k2)[2, "Pr(>Chi)"],
anova(model.0_k2, model.bez_ecog_k2)[2, "Pr(>Chi)"],
anova(model.0_k2, model.bez_karno_k2)[2, "Pr(>Chi)"])
```

Tabela 14: Krok 2: Weryfikacja modeli po usunięciu zmiennej ‘age’

Zmienna usuwana	P-value (IW)
sex	0.0007

Zmienna usuwana	P-value (IW)
ph.ecog	0.0020
ph.karno	0.1756

Jak widzimy w Tabeli 14 zmienna *ph.karno* ma największą wartość poziomu krytycznego ($p - value \approx 0.1756$). Ponieważ wartość ta przekracza przyjęty próg istotności $\alpha = 0.15$, zmienną *ph.karno* możemy uznać za nieistotną i usunąć ją z modelu. Dalej otrzymujemy:

```
#model po usunięciu age i ph.karno
model.0_k3 <- survreg(Surv(time, status) ~ as.factor(sex)
                      + as.factor(ph.ecog),
                      data = dane, dist = "weibull")

#modele zredukowane
model.bez_sex <- survreg(Surv(time, status)
                        ~ as.factor(ph.ecog),
                        data=dane, dist="weibull")

model.bez_ecog <- survreg(Surv(time, status) ~
                        as.factor(sex), data=dane,
                        dist="weibull")

#wartosci p-value
p_value_k3 <- c(
  anova(model.0_k3, model.bez_sex)[2, "Pr(>Chi)"],
  anova(model.0_k3, model.bez_ecog)[2, "Pr(>Chi)"])
```

Tabela 15: Krok 3: Weryfikacja istotności zmiennych w modelu bez zmiennych ‘age’ i ‘ph.karno’

Zmienna w modelu	P-value (IW)
sex	0.00109
ph.ecog	0.00038

W Tabeli 15 sprawdzono istotność zmiennych pozostawionych w modelu (*sex* oraz *ph.ecog*). Wyniki testu ilorazu wiarygodności wskazują, że zmienna *sex* osiągnęła poziom krytyczny $p - value \approx 0.0011$. Zmienna *ph.ecog* osiągnęła poziom krytyczny $p - value \approx 0.0004$. Obie wartości są znacząco mniejsze od przyjętego poziomu istotności ($\alpha = 0.15$). Nie ma zatem podstaw do usunięcia którejkolwiek z pozostałych zmiennych, co oznacza, że obie zmienne są statystycznie istotne i niezbędne w modelu. Procedura eliminacji zostaje zakończona.

Modelem optymalnym w świetle tej metody jest model przyspieszonego czasu awarii zawierający zmienne płeć (*sex*) oraz stopień sprawności (*ph.ecog*). Z modelu początkowego zostały

usunięte zmienne uznane za nieistotne (w metodzie eliminacji wstecznej) takie jak *age* oraz *ph.karno*.

4.3.2 Kod - podpunkt b

W podpunkcie b, mamy dokonać wyboru najlepszego modelu AFT na podstawie kryterium AIC. Jego implementacja opiera się na użyciu funkcji `step`.

```
#model pełny dla AFT
model_pelny <- survreg(Surv(time, status) ~ age_cent
                      + as.factor(sex) + as.factor(ph.ecog)
                      + ph.karno_cent, data = dane, dist = "weibull")

#AIC
model_AFT_AIC <- step(model_pelny, direction = "backward", k = 2)
```

Tabela 16: Historia selekcji modelu (kryterium AIC)

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	218	2250.599	2266.599
- age_cent	1	1.632438	219	2252.231	2266.231
- ph.karno_cent	1	1.834102	220	2254.065	2266.065

Tabela 17: Współczynniki najlepszego modelu wg. AIC

	Value	Std. Error	z	p
(Intercept)	6.1907	0.1244	49.7682	0.0000
as.factor(sex)2	0.3900	0.1240	3.1438	0.0017
as.factor(ph.ecog)1	-0.2964	0.1450	-2.0437	0.0410
as.factor(ph.ecog)2	-0.6789	0.1658	-4.0950	0.0000
as.factor(ph.ecog)3	-1.4200	0.7397	-1.9196	0.0549
Log(scale)	-0.3158	0.0617	-5.1145	0.0000

W Tabeli ?? przedstawiono wyniki procedury automatycznego doboru zmiennych (funkcja `step`), która przeprowadziła dwuetapową redukcję modelu. Początkowa wartość kryterium AIC wynosiła 2266.60. W pierwszym kroku usunięto zmienną *age*, co obniżyło AIC do 2266.23. W drugim kroku usunięto zmienną *ph.karno*, co pozwoliło na dalszy spadek AIC do poziomu 2266.07. Niższa wartość kryterium oznacza lepsze dopasowanie modelu przy uwzględnieniu kary za jego złożoność.

W Tabeli 17 przedstawiono ostateczny model zawierający wyłącznie zmienne *sex* oraz *ph.ecog*. Jest to wynik tożsamy z rezultatem uzyskanym w podpunkcie (a) metodą eliminacji wstecznej. Dodatni współczynnik przy *sex* (0.3900) w modelu AFT oznacza, że bycie kobietą wpływa na wydłużenie czasu przeżycia. Ujemne współczynniki przy *ph.ecog* wskazują, że gorsza sprawność (wyższy numer ECOG) istotnie skraca czas przeżycia.

4.3.3 Kod - podpunkt c

W podpunkcie c, mamy dokonać wyboru najlepszego modelu AFT na podstawie kryterium BIC. Jego implementacja również opiera się na użyciu funkcji `step`.

```
#BIC
n_obs <- sum(dane$status == 1)
model_AFT_BIC <- step(model_pelny, direction = "backward", k=log(n_obs))
```

Tabela 18: Historia selekcji modelu (kryterium BIC)

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	218	2250.599	2291.349
- age_cent	1	1.632438	219	2252.231	2287.887
- ph_karno_cent	1	1.834102	220	2254.065	2284.628

Tabela 19: Współczynniki najlepszego modelu wg. BIC

	Value	Std. Error	z	p
(Intercept)	6.1907	0.1244	49.7682	0.0000
as.factor(sex)2	0.3900	0.1240	3.1438	0.0017
as.factor(ph.ecog)1	-0.2964	0.1450	-2.0437	0.0410
as.factor(ph.ecog)2	-0.6789	0.1658	-4.0950	0.0000
as.factor(ph.ecog)3	-1.4200	0.7397	-1.9196	0.0549
Log(scale)	-0.3158	0.0617	-5.1145	0.0000

W Tabeli 18 widzimy przebieg selekcji - procedure minimalizacji Bayesowskiego Kryterium Informacyjnego (BIC), które nakłada surowszą karę za liczbę parametrów niż AIC, która przebiegała identycznie jak w poprzednich przypadkach.

Początkowa wartość kryterium wynosiła 2291.35. Usunięcie zmiennej *age* obniżyło BIC do 2287.89. Usunięcie zmiennej *ph.karno* pozwoliło osiągnąć minimum na poziomie 2284.63.

Ostateczny model wyłoniony przy użyciu kryterium BIC (Tabela 19) składa się ze zmiennych *sex* oraz *ph.ecog*. Fakt, że wszystkie trzy metody (eliminacja wsteczna oparta na testach, kryterium AIC oraz kryterium BIC) wskazały na ten sam zestaw zmiennych, świadczy o stabilności wyniku i silnej pozycji statystycznej wybranych predyktorów.

Interpretacja współczynników pozostaje bez zmian: płeć żeńska jest czynnikiem protekcyjnym (wydłuża czas przeżycia), natomiast wyższy stopień niesprawności (ECOG) jest czynnikiem ryzyka (skraca czas przeżycia).

4.4 Zadanie 4

Celem zadania jest przeprowadzenie procedury doboru zmiennych (selekcji modelu) analogicznie jak w zadaniu 4.3, tylko w tym przypadku zmieniamy model na model proporcjonalnych

hazardów Coxa (a nie AFT tak jak miało to miejsce w zadaniu ??).

4.4.1 Kod - podpunkt a

Implementacje kodów są analogiczne do tych w widocznych w zadaniu 4.3

```
#model pełny dla proporcjonalnych hazardów Coxa
model_cox_0 <- coxph(Surv(time, status) ~ age_cent
                    + as.factor(sex) + as.factor(ph.ecog)
                    + ph.karno_cent, data = dane)

#modele zredukowane
cox_bez_age <- coxph(Surv(time, status) ~ as.factor(sex)
                    + as.factor(ph.ecog) + ph.karno_cent, data=dane)

cox_bez_sex <- coxph(Surv(time, status) ~ age_cent +
                    as.factor(ph.ecog) + ph.karno_cent, data=dane)

cox_bez_ecog <- coxph(Surv(time, status) ~ age_cent +
                     as.factor(sex) + ph.karno_cent, data=dane)

cox_bez_karno <- coxph(Surv(time, status) ~ age_cent +
                      as.factor(sex) + as.factor(ph.ecog),
                      data=dane)

#obliczanie p-wartości
p_value4_k1 <- c(
  anova(model_cox_0, cox_bez_age)[2, 4],
  anova(model_cox_0, cox_bez_sex)[2, 4],
  anova(model_cox_0, cox_bez_ecog)[2, 4],
  anova(model_cox_0, cox_bez_karno)[2, 4])
```

Tabela 20: Krok 1 (Cox): Weryfikacja wszystkich zmiennych

Zmienna usuwana	P-value (IW)
age	0.1804
sex	0.0006
ph.ecog	0.0036
ph.karno	0.1886

Jak widać w Tabeli 20 zmienna *ph.karno* ma największą wartość poziomu krytycznego ($p - value \approx 0.1886$). Ponieważ wartość ta przekracza przyjęty próg istotności $\alpha = 0.15$, zmienna ta zostaje uznana za nieistotną i usunięta z modelu. Przechodzimy do kroku 2:

```

#model bez ph.karno
model_cox_1 <- coxph(Surv(time, status) ~
                    age_cent + as.factor(sex) + as.factor(ph.ecog),
                    data = dane)

#modele zredukowane
cox_bez_age_k2 <- coxph(Surv(time, status) ~ as.factor(sex)
                      + as.factor(ph.ecog), data=dane)

cox_bez_sex_k2 <- coxph(Surv(time, status) ~ age_cent
                      + as.factor(ph.ecog), data=dane)

cox_bez_ecog_k2 <- coxph(Surv(time, status) ~ age_cent
                      + as.factor(sex), data=dane)

#wartosci p-value
p_value4_k2 <- c(
  anova(model_cox_1, cox_bez_age_k2)[2, 4],
  anova(model_cox_1, cox_bez_sex_k2)[2, 4],
  anova(model_cox_1, cox_bez_ecog_k2)[2, 4])

```

Tabela 21: Krok 2 (Cox): Weryfikacja po usunięciu zmiennej ph.karno

Zmienna usuwana	P-value (IW)
age	0.2343
sex	0.0010
ph.ecog	0.0010

Jak widać w Tabeli ?? zmienna *age* ma największą wartość poziomu krytycznego ($p\text{-value} \approx 0.2343$). Ponieważ wartość ta przekracza przyjęty próg istotności $\alpha = 0.15$, zmienna ta zostaje uznana za nieistotną i usunięta z modelu. Przechodząc dalej otrzymujemy:

```

#model bez ph.karno oraz age
model_cox_2 <- coxph(Surv(time, status) ~ as.factor(sex)
                    + as.factor(ph.ecog), data = dane)

#modele zredukowane
cox_bez_sex_k3 <- coxph(Surv(time, status) ~ as.factor(ph.ecog), data=dane)

cox_bez_ecog_k3 <- coxph(Surv(time, status) ~ as.factor(sex), data=dane)

p_value4_k3 <- c(
  anova(model_cox_2, cox_bez_sex_k3)[2, 4],
  anova(model_cox_2, cox_bez_ecog_k3)[2, 4])

```

Tabela 22: Krok 3 (Cox): Weryfikacja modelu końcowego

Zmienna w modelu	P-value (IW)
sex	0.00103
ph.ecog	0.00035

W kroku 3 (ostatnim etapie analizy) sprawdzono istotność dwóch zmiennych pozostałych w modelu po wcześniejszych eliminacjach. Wyniki testu ilorazu wiarygodności (IW) są jednoznaczne: Dla zmiennej *ph.ecog* wartość p wynosi 0.00035. Dla zmiennej *sex* wartość p -value wynosi 0.00103. Obie wartości są drastycznie niższe od przyjętego poziomu istotności ($\alpha = 0.15$). Oznacza to, że każda z tych zmiennych wnosi istotną informację o ryzyku zgonu i usunięcie którejkolwiek z nich znacząco pogorszyłoby dopasowanie modelu. Procedura eliminacji zostaje zakończona - żadna z pozostałych zmiennych nie może zostać usunięta.

Uzyskany wynik jest identyczny z rezultatem otrzymanym w zadaniu 4.3 dla parametrycznego modelu Weibulla. Świadczy to o dużej odporności wybranych predyktorów – płeć i stan sprawności (ECOG) są kluczowymi determinantami przeżycia w tej grupie pacjentów, niezależnie od przyjętego modelu.

4.4.2 Kod - podpunkt b

```
# Model pełny Coxa
model_cox_pelny <- coxph(Surv(time, status) ~ age_cent
+ as.factor(sex) + as.factor(ph.ecog)
+ ph.karno_cent, data = dane)

#AIC
model_cox_AIC <- step(model_cox_pelny, direction = "backward", k = 2)
```

Tabela 23: Historia selekcji modelu Coxa (AIC)

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	157	1446.537	1458.537
- ph.karno_cent	1	1.728580	158	1448.266	1458.266
- age_cent	1	1.414313	159	1449.680	1457.680

Tabela 24: Współczynniki modelu Coxa wg AIC

	coef	exp(coef)	se(coef)	z	Pr(> z)
as.factor(sex)2	-0.5388	0.5835	0.1684	-3.1997	0.0014
as.factor(ph.ecog)1	0.4186	1.5198	0.1995	2.0987	0.0358
as.factor(ph.ecog)2	0.9315	2.5383	0.2259	4.1239	0.0000
as.factor(ph.ecog)3	2.0689	7.9163	1.0272	2.0141	0.0440

Do wyłonienia optymalnego zestawu zmiennych wykorzystano procedurę eliminacji wstecznej (backward) opartą na kryterium informacyjnym AIC (rezultaty widoczne w Tabeli 23).

Krok 1: Model pełny posiadał AIC na poziomie 1458.537. Krok 2: Usunięcie zmiennej *ph.karno* obniżyło AIC do 1458.266. Krok 3: Usunięcie zmiennej *age* pozwoliło osiągnąć najniższą wartość AIC (1457.680), co skutkowało wyborem modelu z dwiema zmiennymi: *sex* oraz *ph.ecog*.

Model końcowy (Tabela ??) składa się ze zmiennych *sex* oraz *ph.ecog*.

Płeć (*sex*) - ujemny współczynnik regresji (-0.5388) oraz iloraz hazardu ($\exp(\text{coef})$) wynoszący 0.5835 oznaczają, że bycie kobietą zmniejsza ryzyko zgonu. Wynik jest istotny statystycznie ($p = 0.0014$). Stopień sprawności (*ph.ecog*) - wszystkie współczynniki dla poziomów ECOG są dodatnie i rosną wraz z pogorszeniem stanu zdrowia. Wszystkie poziomy zmiennej *ph.ecog* są istotne statystycznie ($p < 0.15$).

Dobór modelu kryterium AIC potwierdził wyniki uzyskane metodą eliminacji wstecznej oraz wyniki otrzymane w zadaniu 4.3. Kluczowymi czynnikami wpływającymi na funkcję hazardu są płeć (czynnik chroniący) oraz stopień niesprawności (czynnik ryzyka).

4.4.3 Kod - podpunkt c

```
#BIC
model_cox_BIC <- step(model_cox_pelny, direction = "backward", k = log(n_obs))
```

Tabela 25: Historia selekcji modelu Coxa (BIC)

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	157	1446.537	1477.100
- ph_karno_cent	1	1.728580	158	1448.266	1473.734
- age_cent	1	1.414313	159	1449.680	1470.055

Tabela 26: Współczynniki modelu Coxa wg BIC

	coef	$\exp(\text{coef})$	se(coef)	z	$\Pr(> z)$
as.factor(sex)2	-0.5388	0.5835	0.1684	-3.1997	0.0014
as.factor(ph.ecog)1	0.4186	1.5198	0.1995	2.0987	0.0358
as.factor(ph.ecog)2	0.9315	2.5383	0.2259	4.1239	0.0000
as.factor(ph.ecog)3	2.0689	7.9163	1.0272	2.0141	0.0440

Procedura minimalizacji Bayesowskiego Kryterium Informacyjnego (BIC) (Tabela 25), przebiegła według identycznego schematu co AIC. Wartość wyjściowa kryterium dla modelu pełnego wynosiła 1477.100. W pierwszym kroku usunięto zmienną *ph.karno*, redukując BIC do 1473.734. W drugim kroku usunięto zmienną *age*, osiągając minimum na poziomie 1470.055.

Niższa wartość BIC wskazuje na model, który jest bardziej prawdopodobny w świetle danych, przy uwzględnieniu kary za liczbę parametrów.

Ostateczny model wyłoniony metodą BIC (Tabela 26) zawiera zmienne *sex* oraz *ph.ecog*.

Wszystkie trzy zastosowane metody doboru zmiennych (manualna eliminacja wsteczna, AIC oraz BIC) dla modelu Coxa doprowadziły do tego samego zestawu predyktorów. Potwierdza to silną i stabilną relację: płeć żeńska jest czynnikiem istotnie obniżającym ryzyko zgonu (współczynnik ujemny), natomiast każdy kolejny stopień niesprawności w skali ECOG istotnie zwiększa to ryzyko (współczynniki dodatnie).