Sprawozdanie 3

Analiza przeżycia

Marta Stankiewicz (282244) Kacper Szmigielski (282255)

Spis treści

Spis rysunków

Spis tabel

1 Lista 9

Sprawozdanie dotyczy analizy zbioru *lung* dostępnych w pakiecie survival. Dane dotyczą pacjentów z zaawansowanym rakiem płuc. Zostały one odpowiednio przygotowane, tzn. pominięto zmienne niepotrzebne dla analizy (kolumny: *inst*, *pat.karno*, *meal.cal*, *wt.loss*), usunięto wiersze zawierające wartości brakujące oraz wycentrowano zmienne ciągłe (poprzez odjęcie średniej wartości). W poniższych zadaniach przyjęto, że czas przeżycia ma rozkład Weibulla.

1.1 Zadanie 1

W pierwszej kolejności oszacowano (zgodnie z przykładem zawartym na stronie 8 wykładu 9) parametry modelu przyspieszonego czasu awarii, przyjmując za zmienną zależną zmienną *time*, a za charakterystyki zmienne: *age*, *sex*, *ph.ecog*, *ph.karno*.

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


beta <- -summary(model)$coefficients[-1]
mu <- model$icoef[1]
sigma <- exp(model$icoef[2])
alpha <-  1/sigma
lambda <- exp(-mu*alpha)

wsp.AFT <- data.frame(
  "alpha" = alpha,
  "beta" = beta,
  "lambda" = lambda,
  "mu" = mu,
  "sigma" = sigma
)

rownames(wsp.AFT) <- c("age",
                       "sex",
                       "ph.ecog = 1",
                       "ph.ecog = 2",
                       "ph.ecog= 3",
                       "ph.karno")
```

1.2 Zadanie 2

Tabela 1: Współczynniki w modelu AFT

Zmienne	α	β	λ	μ	σ
age	1.32569	0.00860	0.00033	6.0431	0.75432
sex	1.32569	-0.40828	0.00033	6.0431	0.75432
ph.ecog = 1	1.32569	0.42156	0.00033	6.0431	0.75432
ph.ecog = 2	1.32569	0.92611	0.00033	6.0431	0.75432
ph.ecog= 3	1.32569	1.68724	0.00033	6.0431	0.75432

Zmienne	α	β	λ	μ	σ
ph.karno	1.32569	0.01006	0.00033	6.0431	0.75432

Analizując otrzymane współczynniki widoczne w tabeli 1 można zauważyć, że jedyne zmiany widoczne są wśród parametrów β. Wykorzystywana w pakiecie R funkcja *survreg* estymuje parametry modelu AFT, który w postaci liniowej wyraża się wzorem:

$$ln(X_z) = \mu - \beta^T z + \sigma W$$

gdzie $\beta^T z = \beta_1 z_1 + ... + \beta_n z_n$ jest iloczynem skalarnym wektora współczynników i zmiennych objaśniających. Pozostałe elementy oznaczają:

μ - przesunięcie (poziom log-czasu, gdy z = 0)

σW - błąd losowy (postać rozkładu zmiennej W jest znana)

Warto zaznaczyć, że parametry α oraz λ, charakterystyczne dla postaci ogólnej rozkładów, są funkcjami μ oraz σ i pojawiają się dopiero po przekształceniu modelu do formy hazardu lub funkcji przeżycia. Z tego względu w bezpośredniej interpretacji wpływu zmiennych objaśniających kluczowe znaczenie mają współczynniki β.

Jeśli $\beta > 0$, to wartość tę odejmujemy w równaniu, co prowadzi do zmniejszenia logarytmu czasu, a w konsekwencji do skrócenia przewidywanego czasu przeżycia pacjenta.Jeśli $\beta < 0$, wpływ zmiennej jest odwrotny – prowadzi to do zwiększenia wartości $ln(X_z)$, co interpretujemy jako wydłużenie oczekiwanego czasu przeżycia.

Tabela 2: Współczynniki β w modelu AFT

Zmienne	β
age	0.00860
sex	-0.40828
ph.ecog = 1	0.42156
ph.ecog = 2	0.92611
ph.ecog= 3	1.68724
ph.karno	0.01006

Z analizy parametrów zamieszczonych w tabeli 2 wynika, że jedyną zmienną o ujemnym współczynniku β jest *sex*. Ponieważ jest to zmienna typu faktor, sprawdzany jest jej poziom referencyjny.

```
levels(as.factor(lung$sex))
```

```
## [1] "1" "2"
```

Kategorią odniesienia (kodowaną jako 1) są mężczyźni. Ujemna wartość współczynnika dla poziomu 2 (kobiety) oznacza, że w przyjętej konwencji modelu — gdzie dodatnie β skraca czas przeżycia — płeć żeńska jest czynnikiem sprzyjającym wydłużeniu oczekiwanego czasu przeżycia w stosunku do mężczyzn.

Kolejną istotną zmienną jakościową jest ph.ecog, określająca stan sprawności pacjenta w skali ECOG.

```
levels(as.factor(lung$ph.ecog))
```

```
## [1] "0" "1" "2" "3"
```

W tym przypadku poziomem referencyjnym jest wartość 0 (pacjent w pełni sprawny). Dodatnie wartości współczynników β dla poziomów 1, 2 oraz 3 wskazują, że każdy stopień pogorszenia sprawności fizycznej wiąże się ze skróceniem przewidywanego czasu przeżycia w porównaniu do grupy o najwyższej sprawności (poziom 0).

Pozostałe zmienne w modelu mają charakter ciągły. Ich dodatnie współczynniki β sugerują, że wzrost wartości tych parametrów skraca czas przeżycia.

1.3 Zadanie 3

W modelu przyspieszonego czasu życia (AFT) zakłada się, że funkcja przeżycia jednostki o wektorze charakterystyk z może zostać zapisana jako przeskalowanie czasu w bazowej funkcji przeżycia:

$$S_z(t) = S_0\Big(\exp\Big(\beta^T z\Big)\ t\Big)\,, \tag{1}$$

gdzie $S_0(t)$ oznacza funkcję przeżycia jednostki bazowej, tj. odpowiadającej zerowemu wektorowi charakterystyk.

Jeżeli bazowy czas przeżycia ma rozkład Weibulla z parametrami $\lambda > 0$ oraz $\alpha > 0$, to bazowa funkcja przeżycia dana jest wzorem

$$S_0(t) = \exp(-\lambda\, t^{\alpha})\,, \qquad t \geq 0. \tag{2}$$

Podstawiając (2) do definicji (1), otrzymujemy funkcję przeżycia jednostki o wektorze charakterystyk z w postaci

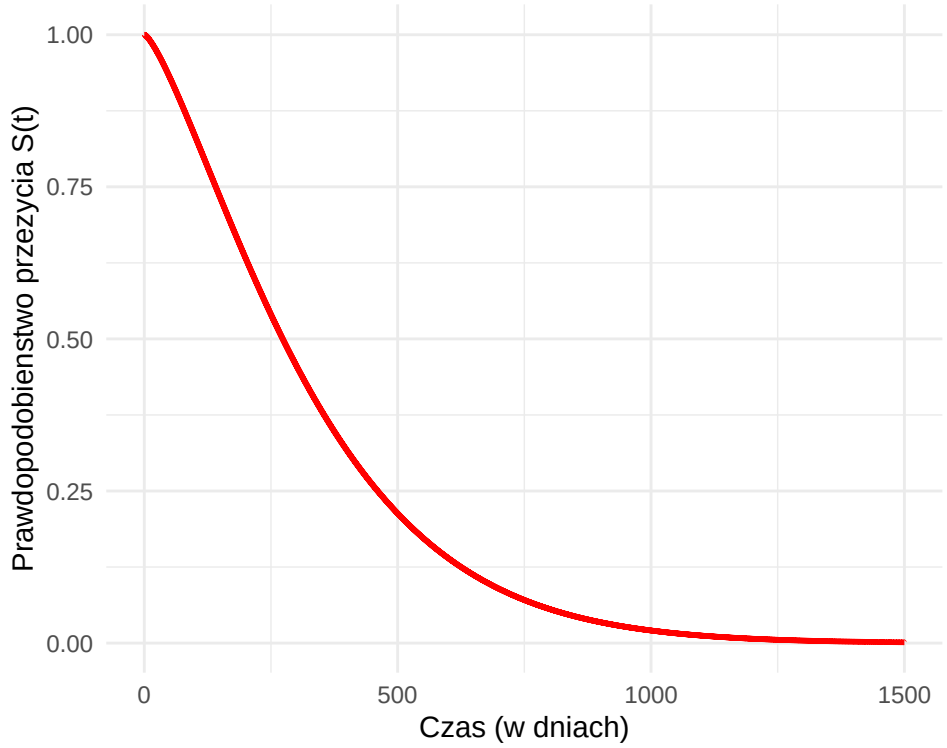
$$S_z(t) = \exp\Big(-\lambda\, \exp\Big(\alpha\, \beta^{\top} z\Big)\ t^{\alpha}\Big)\,. \tag{3}$$

```
pacjent <- c(70 - mean_wiek, 1, 1, 0, 0, 90 - mean_ph_karno)

S_0 <- function(x) exp(-lambda*(x^alpha))

S <- function (t) {
  exp(-lambda*exp(alpha*sum(beta*pacjent))*t^alpha)
}
```

Wykorzystując wyprowadzony wzór na funkcję przeżycia, wyznaczono prawdopodobieństwo przeżycia powyżej 300 dni dla 70-letniej pacjentki o charakterystyce $ph.ecog = 1$ oraz $ph.karno = 90$. Prawdopodobieństwo to wynosi 0.4554.

1.4 Zadanie 4

Poniżej przedstawiono wykres oszacowanej w zadaniu 3. funkcji przeżycia.



Rysunek 1: Wykres oszacowanej funkcji przeżycia - model AFT

1.5 Zadanie 5

Definicja 1.1. *Niech $h_0(x)$ będzie funkcją hazardu (o znanej postaci) obserwowalnej zmiennej losowej X, odpowiadającej jednostce o zerowym wektorze charakterystyk. Wówczas model, w którym funkcja hazardu jednostki o wektorze charakterystyk z ma postać*

$$h_z(x) = h_0(x)\, \exp\Big(\beta^T z\Big)\,, \tag{4}$$

nazywamy (parametrycznym) modelem proporcjonalnych hazardów (PH). Funkcję $h_0(x)$ nazywamy bazową funkcją hazardu.

Model proporcjonalnych hazardów charakteryzuje się prostą własnością interpretacyjną. Jeżeli z_1 oraz z_2 są dwoma wektorami charakterystyk, to na podstawie (4) otrzymujemy

$$\frac{hz_1(x)}{hz_2(x)} = \exp\Big(\beta^\top (z_1 - z_2)\Big)\,, \tag{5}$$

co oznacza, że iloraz hazardów dwóch jednostek jest stały w czasie, a zatem hazardy są proporcjonalne.

```
model <-  phreg(
  Surv(time, status)~age + as.factor(sex) + as.factor(ph.ecog) + ph.karno,
  data = dane,
  dist = "weibull"
  )

beta <- model$coefficients[-c(7,8)]
mu <- model$coefficients['log(scale)']
sigma <- exp(model$coefficients['log(shape)'])
lambda <- exp(-mu*sigma)
alpha <- sigma
```

1.6 Zadanie 6

Tabela 3: Współczynniki w modelu PH

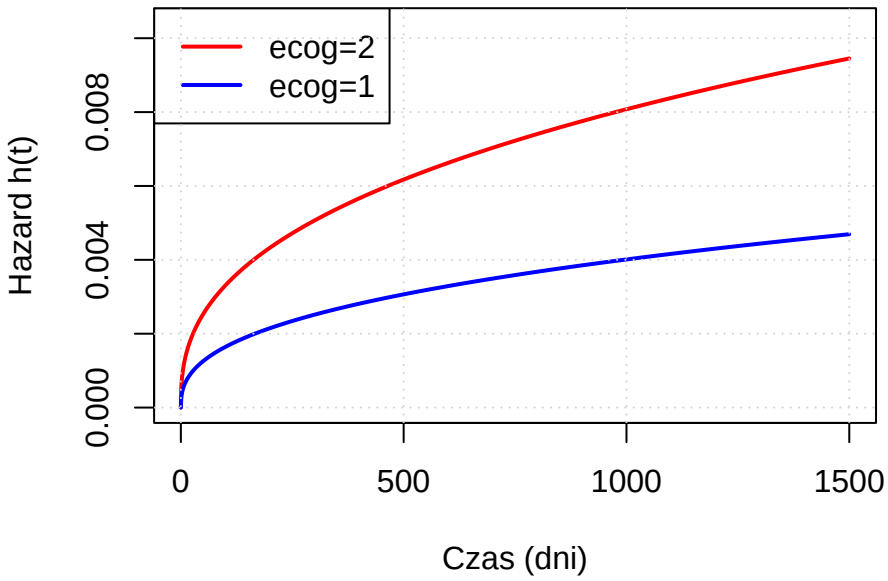
Zmienne	α	β	λ	μ	σ
age	1.38763	0.01194	0.00016	6.30118	1.38763
sex	1.38763	-0.56654	0.00016	6.30118	1.38763
ph.ecog = 1	1.38763	0.58497	0.00016	6.30118	1.38763
ph.ecog = 2	1.38763	1.28510	0.00016	6.30118	1.38763
ph.ecog= 3	1.38763	2.34127	0.00016	6.30118	1.38763
ph.karno	1.38763	0.01395	0.00016	6.30118	1.38763

Obserwując wyniki w tabeli 3 jedynym zidentyfikowanym czynnikiem redukującym ryzyko jest płeć żeńska, dla której współczynnik β wynosi -0.566542. Na jego podstawie można obliczyć wartość $\exp(\beta) \approx 0.567$, oznaczający ryzyko zgonu niższe o 43.3% w stosunku do mężczyzn. W przypadku zmiennej jakościowej *ph.ecog*, dodatnie wartości współczynników β dla poziomów 1, 2 oraz 3 (wynoszące odpowiednio: 0.5849707, 1.2851033 oraz 2.3412704 wskazują na wzrost ryzyka wraz z pogarszającym się stanem sprawności pacjenta — dla $ph.ecog = 3$ hazard jest aż 10.39-krotnie wyższy niż w grupie referencyjnej. Pozostałe zmienne ciągłe, *age* oraz *ph.karno*, również charakteryzują się dodatnimi wartościami β (odpowiednio 0.0119361 oraz 0.013955), co interpretujemy jako wzrost hazardu wraz ze zwiększaniem się wartości tych parametrów, co w konsekwencji prowadzi do skrócenia prognozowanego czasu przeżycia.

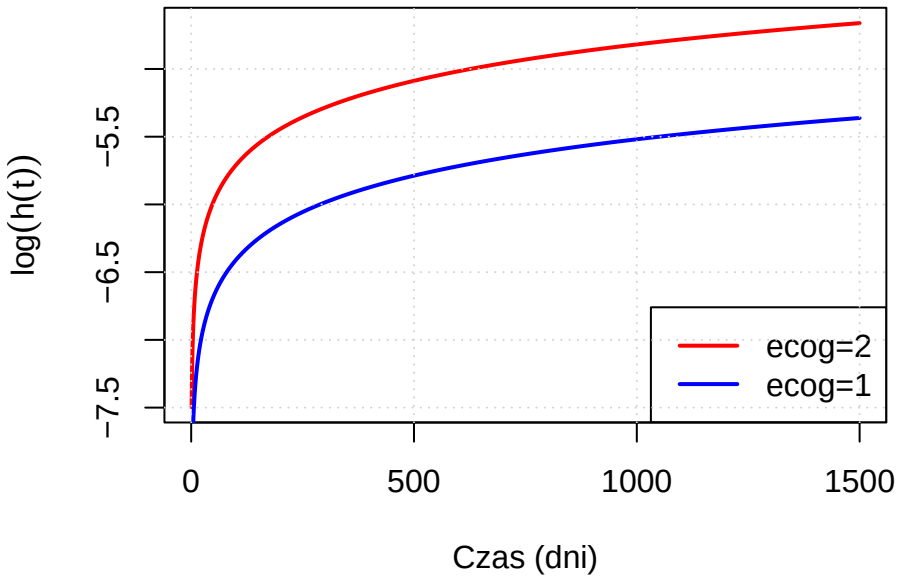
1.7 Zadanie 7

Na podstawie dopasowanego modelu PH wyznaczono oszacowania funkcji hazardu dla 70-letniej kobiety ($ph.karno = 90$), uwzględniając dwa alternatywne poziomy sprawności: $ph.ecog = 1$ oraz $ph.ecog = 2$.

```r
z1 <- c(70 - mean_wiek, 1, 1, 0, 0, 90 - mean_ph_karno)
z2 <- c(70 - mean_wiek, 1, 0, 1, 0, 90 - mean_ph_karno)

h_0 <- function(x) {
  return (lambda*alpha*x^(alpha-1))
  }

h <- function(x, z) {
  h_0(x)*exp(sum(beta*z))
}

x <- 0:1500

haz1 <- sapply(x, function(t) h(t, z1))
haz2 <- sapply(x, function(t) h(t, z2))
```



Rysunek 2: Porównanie estymowanych funkcji hazardu dla różnych poziomów sprawności ECOG - model PH



Rysunek 3: Porównanie logarytmicznych funkcji hazardu dla wybranych poziomów sprawności ECOG - model PH

Weryfikacja założeń modelu, przeprowadzona poprzez analizę wyników przedstawionych na wykresach 2 i 3, potwierdza słuszność przyjęcia modelu proporcjonalnych hazardów. Stały odstęp między krzywymi $\log(h(t))$ dla różnych poziomów zmiennej *ph.ecog* dowodzi, że iloraz hazardu pozostaje stały w czasie.

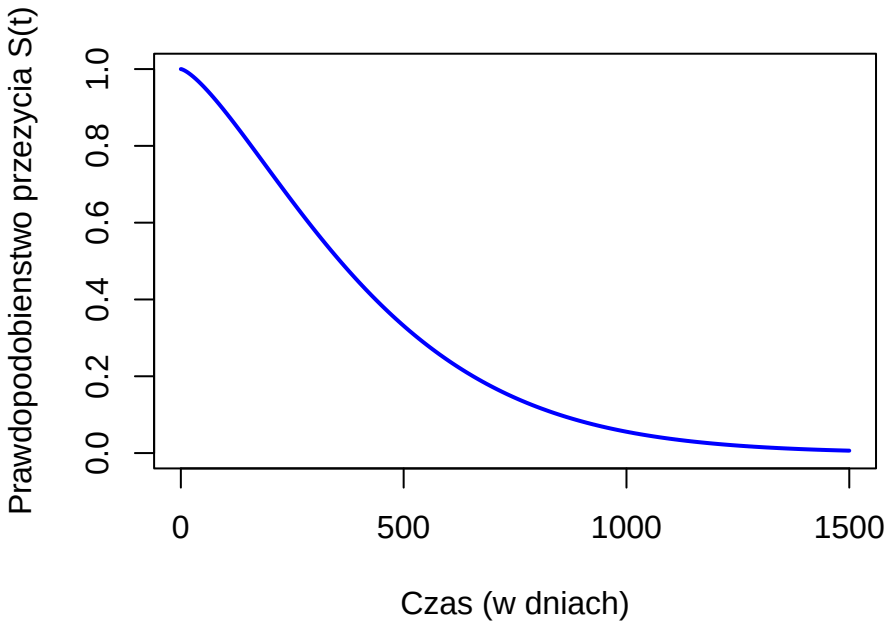
1.8 Zadanie 8

W modelu proporcjonalnych hazardów (PH) funkcja przeżycia dla jednostki o wektorze cech z jest definiowana jako bazowa funkcja przeżycia podniesiona do potęgi odpowiadającej ilorazowi hazardu. Przyjmuje ona następującą postać:

```
S_0 <- function(x) exp(-lambda*(x^alpha))

S <- function (t, z) {
  (S_0(t))^exp(sum(beta*z))
}
```

Prawdopodobieństwo, że czas życia kobiety w wieku 70 lat o charakterystyce $ph.ecog = 1$ i $ph.karno = 90$ będzie większy niż 300 dni wynosi 0.58061. Natomiast prawdopodobieństwo, że czas życia kobiety w wieku 70 lat o charakterystyce $ph.ecog = 2$ i $ph.karno = 90$ będzie większy niż 300 dni wynosi 0.33455. Porównując wynik dla kobiety o charakterystykach opisanych w punkcie (a) z tym otrzymanym w zadaniu 3 możemy zauważyć, że oszacowane prawdopodobieństwo jest większe, gdy korzystamy z modelu PH zamiast AFT.

1.9 Zadanie 9



Rysunek 4: Wykres oszacowanej funkcji przeżycia - model PH

Porównując wykresy 1 oraz 4 można zauważyć różnicę w tempie spadku oszacowanej funkcji przeżycia. Krzywa odpowiadająca modelowi proporcjonalnych hazardów (PH) maleje wolniej niż w przypadku modelu przyspieszonego czasu życia (AFT). W rezultacie dla tych samych punktów czasowych model PH generuje systematycznie wyższe prawdopodobieństwa przeżycia, co skutkuje uzyskaniem bardziej optymistycznych prognoz w porównaniu do modelu AFT.

2 Lista 10

W poniższych z adaniach, nie przyjęto żadnego konkretnego rozkładu czasu życia.

2.1 Zadanie 1

W modelu proporcjonalnych hazardów Coxa zakłada się, że rozkład czasu do wystąpienia zdarzenia jednostki o charakterystyce z ma funkcję hazardu postaci

$$h_z(t) = h_0(t)\psi(z)$$

Najczęściej przyjmuje się, że

$$\psi(z) = exp(\beta^T z)$$

Zatem

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

Poniżej oszacowano parametry modelu proporcjonalnych hazardów Coxa, przyjmując za zmienną zależną zmienną *time*, a za charakterystyki zmienne: *age*, *sex*, *ph.ecog* oraz *ph.karno*.

```
model <- coxph(Surv(time, status) ~ age + factor(sex) + factor(ph.ecog) + ph.karno,
               data = dane, ties = "efron")


s <- summary(model)

beta  <- s$coefficients[, "coef"]
pval  <- s$coefficients[, "Pr(>|z|)"]
HR    <- s$coefficients[, "exp(coef)"]

# 95% CI dla HR:
CI <- s$conf.int[, c("lower .95", "upper .95")]

wyniki <- cbind(beta=beta, HR=HR, CI, p.value=pval)
```

2.2 Zadanie 2

Tabela 4: Wartości oszacowanych parametrów wraz z ilorazami hazardu, przedziałami ufności oraz p-wartościami - model Coxa

	beta	HR	lower .95	upper .95	p.value
age	0.01256	1.01264	0.99405	1.03157	0.18390
factor(sex)2	-0.56568	0.56798	0.40723	0.79218	0.00086
factor(ph.ecog)1	0.57806	1.78257	1.12157	2.83316	0.01447
factor(ph.ecog)2	1.23990	3.45525	1.72205	6.93289	0.00048
factor(ph.ecog)3	2.39585	10.97756	1.31878	91.37750	0.02670
ph.karno	0.01242	1.01250	0.99365	1.03171	0.19512

Analizując oszacowania współczynników β wraz z ich ilorazami hazardu, przedziałami ufności oraz p-wartościami przedstawionymi w tabeli 4, można zauważyć wyraźny podział na zmienne istotne i nieistotne statystycznie. Zmienne *age* oraz *ph.karno* nie wykazują istotnego wpływu na ryzyko zgonu przy przyjętym poziomie istotności $\alpha = 0.05$, ponieważ ich p-wartości wynoszą odpowiednio 0.1839 oraz 0.1951 - są większe od przyjętego poziomu 0.05.

Pozostałe parametry modelu są istotne statystycznie. Dla zmiennej *sex* współczynnik β przyjmuje wartość ujemną (-0.5657), co przekłada się na iloraz hazardu $HR = \exp(\beta) \approx 0.568$. Oznacza to redukcję ryzyka zgonu o około 43.2% w porównaniu do mężczyzn. Z kolei w przypadku zmiennej *ph.ecog* obserwujemy dodatnie wartości współczynników, które rosną wraz z pogarszającym się stanem pacjenta. Dla najwyższego poziomu niesprawności (*ph.ecog* = 3) iloraz hazardu wynosi 10.98, co wskazuje na wielokrotny wzrost ryzyka względem grupy referencyjnej, przy czym szeroki przedział ufności dla tej kategorii sugeruje mniejszą precyzję oszacowania.

2.3 Zadanie 3

Poniżej wyznaczono oszacowanie bazowej skumulowanej funkcji hazardu i bazowej funkcji przeżycia odpowiadającej rozkładowi czasu życia.

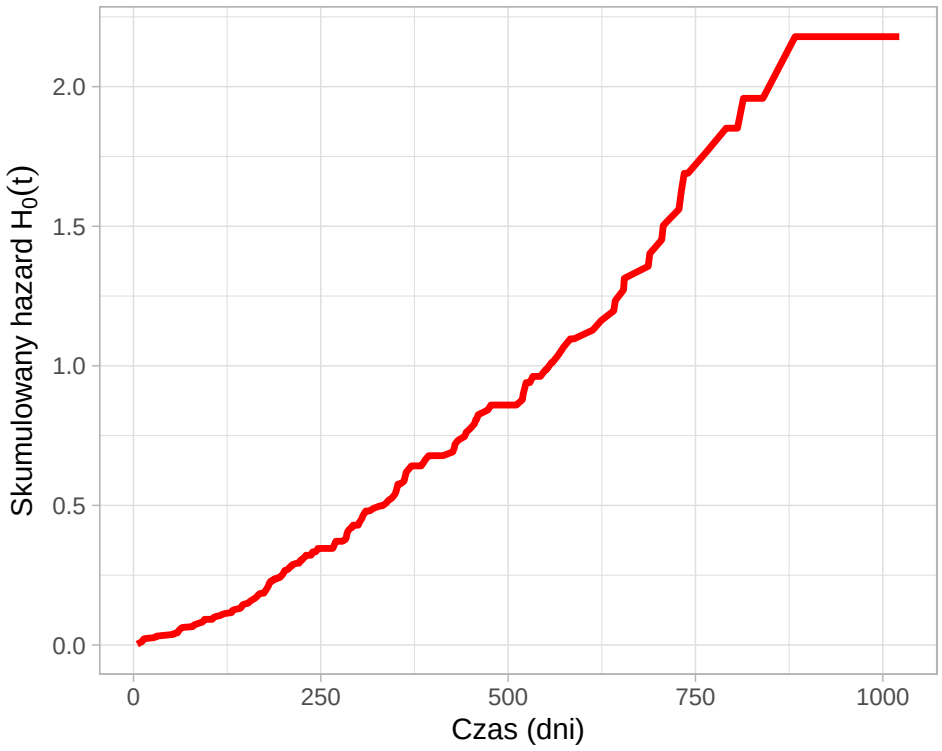
Wykorzystano poniższe wzory.

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

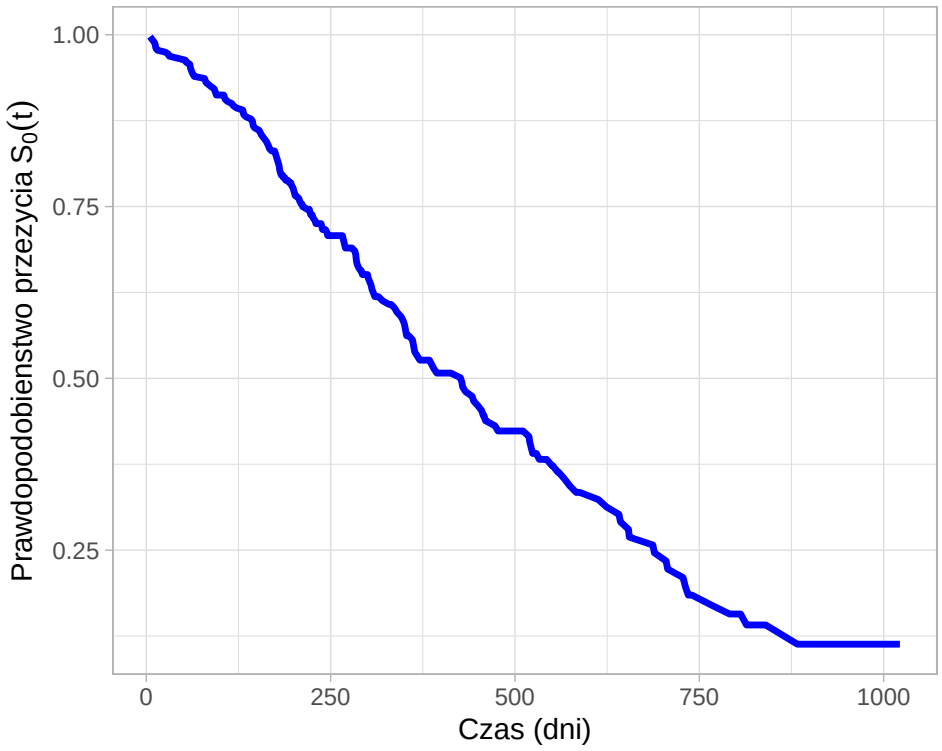
$$H_z(t) = H_0(t) \exp(\beta^T z)$$

$$S_z(t) = \Big(S_0(t)\Big)^{\exp(\beta^T z)}$$

$$S_z(t) = \exp\Big(- H_z(t)\Big)$$



Rysunek 5: Oszacowanie bazowej skumulowanej funkcji hazardu odpowiadającej rozkładowi czasu życia - model Coxa



Rysunek 6: Oszacowanie bazowej funkcji przeżycia odpowiadającej rozkładowi czasu życia - model Coxa

2.4 Zadanie 4

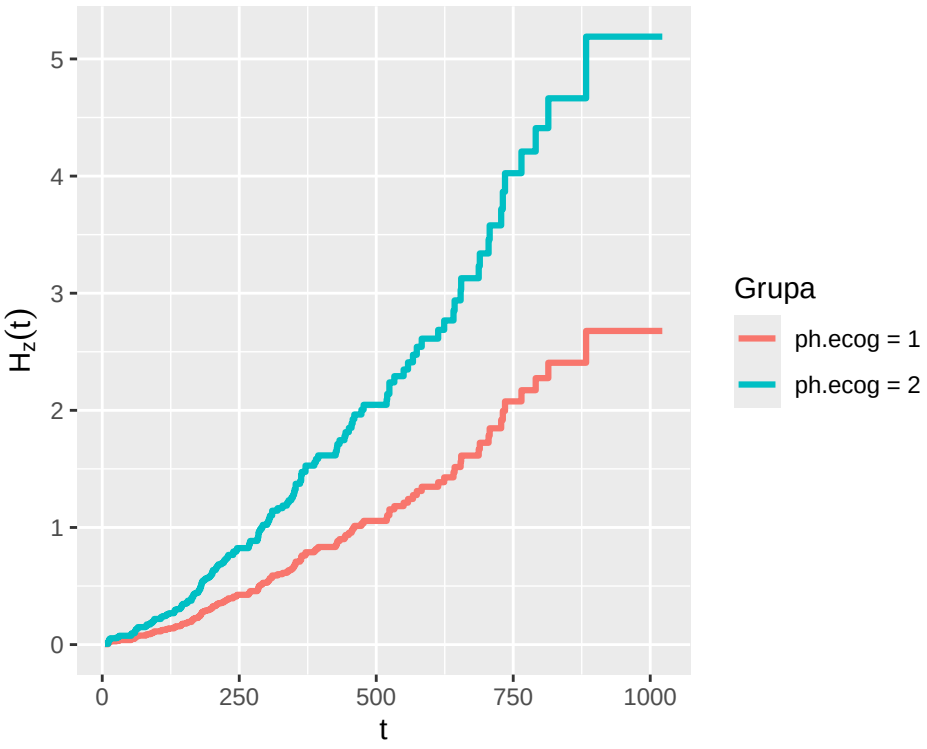
Dla modelu Coxa skumulowana funkcja hazardu dla pacjenta o wektorze cech z wyraża się wzorem:

$$H_z(t) = H_0(t) \exp\Big(\beta^T z\Big)$$

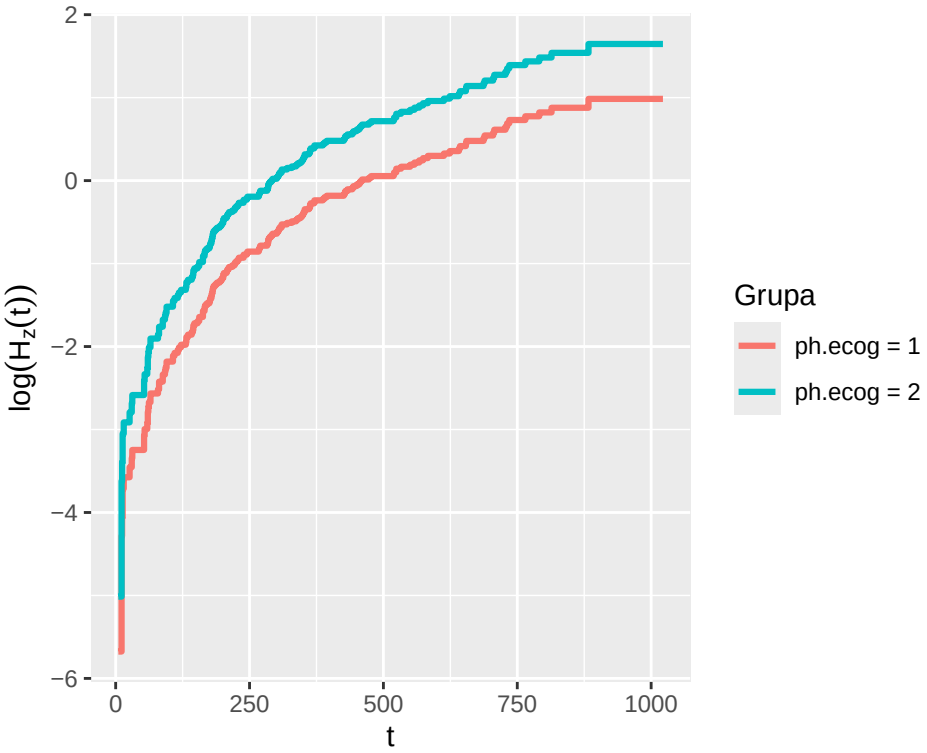
gdzie $H_0(t)$ to bazowy skumulowany hazard, a czynnik $exp(\beta^T z)$ jest stały w czasie dla ustalonego pacjenta. Logarytmując powyższe równanie stronami otrzymujemy:

$$\log H_z(t) = \log H_0(t) + \beta^T z$$

Oznacza to, że na wykresie logarytmicznym krzywe dla różnych grup powinny być przesunięte względem siebie o stałą wartość, a odległość między nimi odpowiada różnicy w ryzyku $\beta^T(z_1 - z_2)$.



Rysunek 7: Oszacowanie skumulowanej funkcji hazardu - model Coxa



Rysunek 8: Logarytm skumulowanej funkcji hazardu - model Coxa

Analiza wykresów 7 oraz 8 potwierdza poprawność zastosowania modelu proporcjonalnych hazardów dla badanych danych. Na pierwszym wykresie widać, że krzywa dla pacjentki z gorszym rokowaniem ($ph.ecog = 2$) jest proporcjonalnie przeskalowana w górę względem krzywej referencyjnej, zachowując podobny kształt przebiegu. Wniosek ten jest jeszcze wyraźniejszy na drugim wykresie przedstawiającym logarytmy funkcji. Krzywe $\ln(H_z(t))$ są względem siebie równolegle przesunięte, zachowując stały dystans w całym analizowanym przedziale czasowym. Zatem iloraz hazardu jest stały w czasie, co oznacza, że założenia modelu Coxa są spełnione i nie ma podstaw do ich kwestionowania.

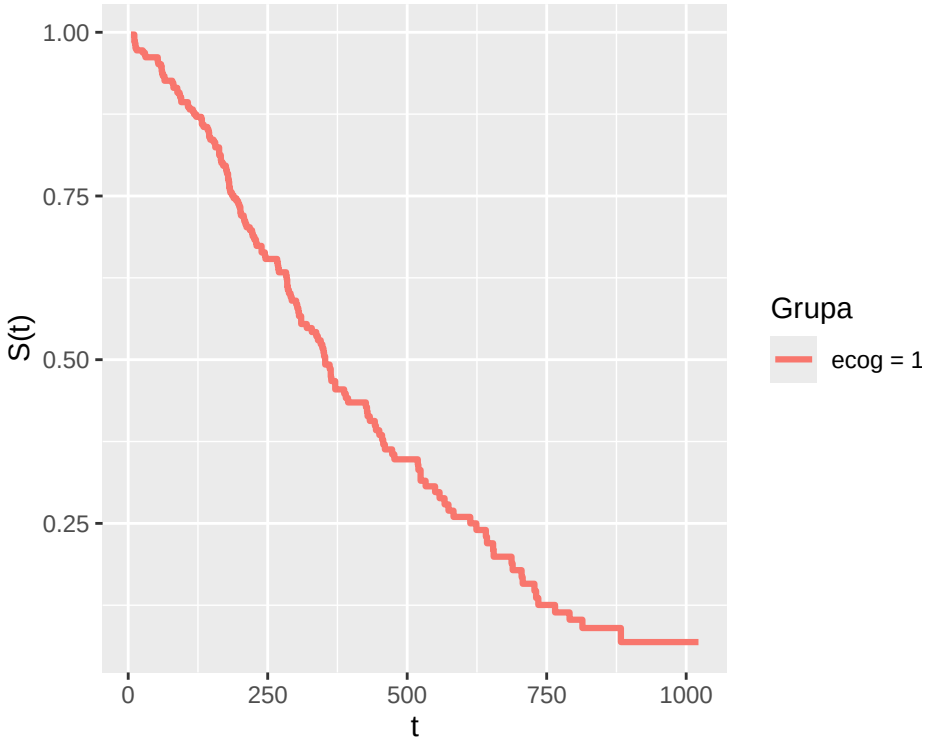
2.5 Zadanie 5

Wyznaczono oszacowanie funkcji przeżycia (w dniach) odpowiadającej rozkładowi czasu życia kobiet w wieku 70 lat ($ph.karno = 90$), których $ph.ecog = 1$ lub $ph.ecog = 2$:

```
nd <- data.frame(
  age = c(70, 70)-mean_wiek,
  sex = factor(c(2, 2), levels = levels(factor(dane$sex))),
  ph.ecog = factor(c(1, 2), levels = levels(factor(dane$ph.ecog))),
  ph.karno = c(90, 90)-mean_ph_karno
)

fit <- survfit(model, newdata = nd)

wynik <- summary(fit, times = 300)

prob_a <- wynik$surv[1]
prob_b <- wynik$surv[2]
```

Analiza uzyskanych wyników wskazuje na istotne różnice w rokowaniach zależne od stanu sprawności. Szacowane prawdopodobieństwo, że czas życia kobiety o charakterystyce $ph.ecog = 1$ (i $ph.karno = 90$) przekroczy 300 dni, wynosi 0.5901. Wartość ta jest większa niż te otrzymane dla modeli AFT oraz PH. W przypadku pacjentki o gorszym stanie sprawności ($ph.ecog = 2$), prawdopodobieństwo ulega wyraźnemu obniżeniu i wynosi 0.3598 dla modelu Coxa. Jest to wartość większa niż wynik otrzymany dla modelu PH.

2.6 Zadanie 6



Rysunek 9: Oszacowanie funkcji przeżycia - model Coxa

Zestawienie funkcji przeżycia na wykresach 4 i 9 wykazuje różnicę w charakterze estymacji: model Coxa generuje funkcję schodkową, zmieniającą się wyłącznie w momentach wystąpienia zdarzeń, co wynika z braku założeń co do kształtu hazardu bazowego. Z kolei model parametryczny PH prezentuje gładką aproksymację tego procesu, wynikającą z przyjęcia teoretycznego rozkładu czasu życia - rozkładu Weibulla. Zbliżony przebieg obu krzywych (trend spadkowy) świadczyłby o tym, że przyjęty rozkład parametryczny dobrze odzwierciedla rzeczywistą strukturę danych empirycznych.

3 Lista 11

3.1 Zadanie 1

Dopasowanie modelu (oszacowanie parametrów)

```
model<-prop.odds(Event(time, cause = status) ~ age +
                 factor(sex) +
                 factor(ph.ecog) +
                 ph.karno,
                 data = dane,n.sim=500,profile=1)
```

3.2 Zadanie 2

Korzystając z poniższych wzorów możemy podać interpretację:

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

$$\ln\left(\frac{\theta_{z_1}(t)}{\theta_{z_2}(t)}\right) = \beta^T(z_1 - z_2)$$
$$\ln\left(\theta_{z_1}(t)\right) = \beta^T(z_1).$$

Tabela 5: Tabela otrzymanych β dla modelu proporcjonalnych szans

Zmienne	β
age	0.00238
factor(sex)2	-0.56145
factor(ph.ecog)1	0.31959
factor(ph.ecog)2	1.14847
factor(ph.ecog)3	1.85058
ph.karno	-0.00225

- Zmienna `age` ma współczynnik β = 0.002378>0, co oznacza, że wraz ze wzrostem wieku rosną szanse (odds) wystąpienia zdarzenia do czasu t (przy stałych pozostałych zmiennych). Równoważnie:
$$\exp(\beta) > 1$$

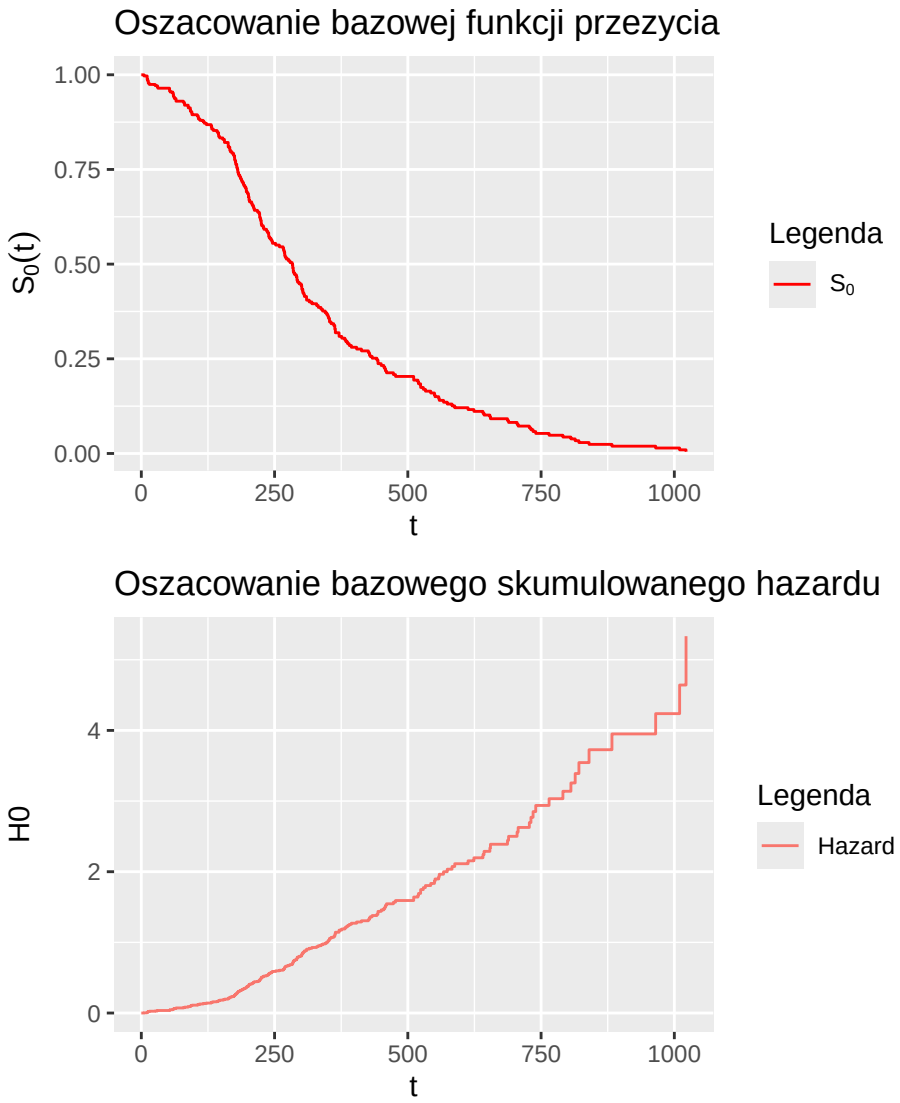
- Zmienna `sex` (kategoryczna) ma współczynnik β = -0.5614481>0 dla poziomu porównywanego do poziomu referencyjnego (bazowego). Oznacza to, że w tej grupie szanse (odds) zajścia zdarzenia do czasu t są mniejsze niż w grupie bazowej, tj.
$$\exp(\beta) < 1$$

- Zmienna `ph.ecog` (kategoryczna) ma współczynniki $\beta > 0$ dla poszczególnych poziomów w porównaniu do poziomu referencyjnego (np. 1 vs 0, 2 vs 0, 3 vs 0). Wskazuje to, że osoby z wyższym `ph.ecog` mają większe szanse (odds) wystąpienia zdarzenia do czasu t niż osoby z poziomu bazowego, czyli
$$\exp(\beta) > 1$$

- Dla zmiennej ciągłej `ph.karno` otrzymano β = -0.0022452<0 co oznacza, że wraz ze wzrostem `ph.karno` maleją szanse (odds) zajścia zdarzenia do czasu t, tj.
$$\exp(\beta) < 1$$

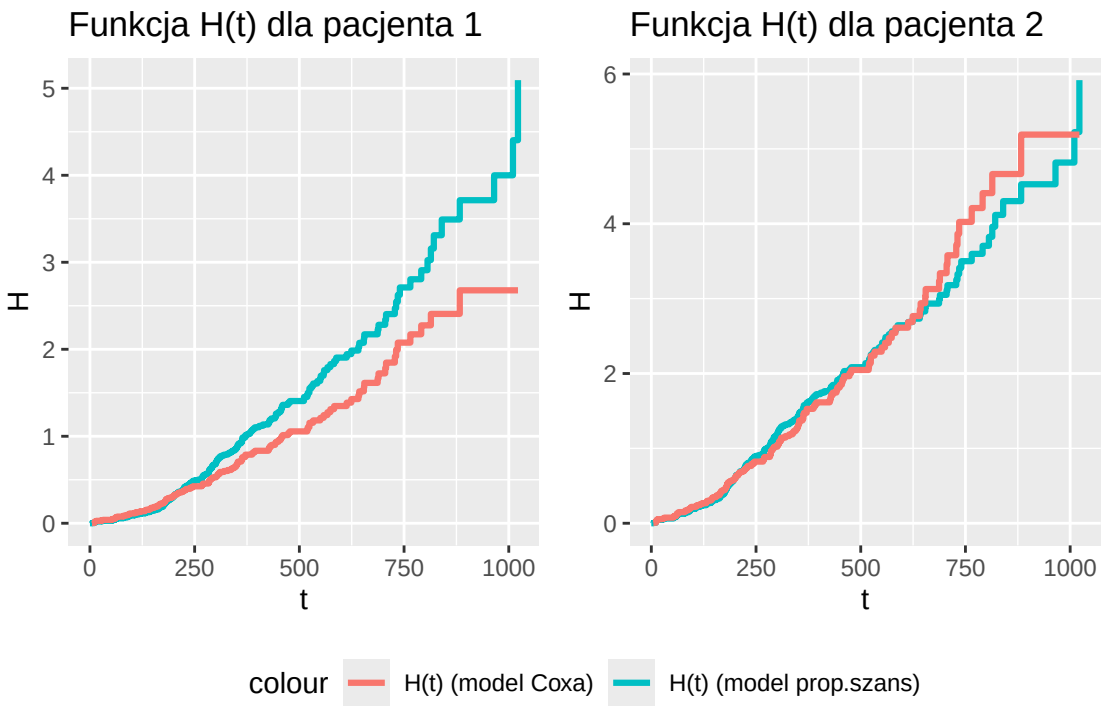
Jeśli jednak w poprzedniej analizie (Lista 10) współczynnik dla `ph.karno` był nieistotny statystycznie, to również tutaj wniosek o kierunku efektu należy traktować ostrożnie: przy braku istotności znak współczynnika może zmieniać się między dopasowaniami/modelami i nie powinien być interpretowany jako stabilny efekt.

3.3 Zadanie 3



Rysunek 10: Bazowa funkcja przeżycia i skumulowanego hazardu

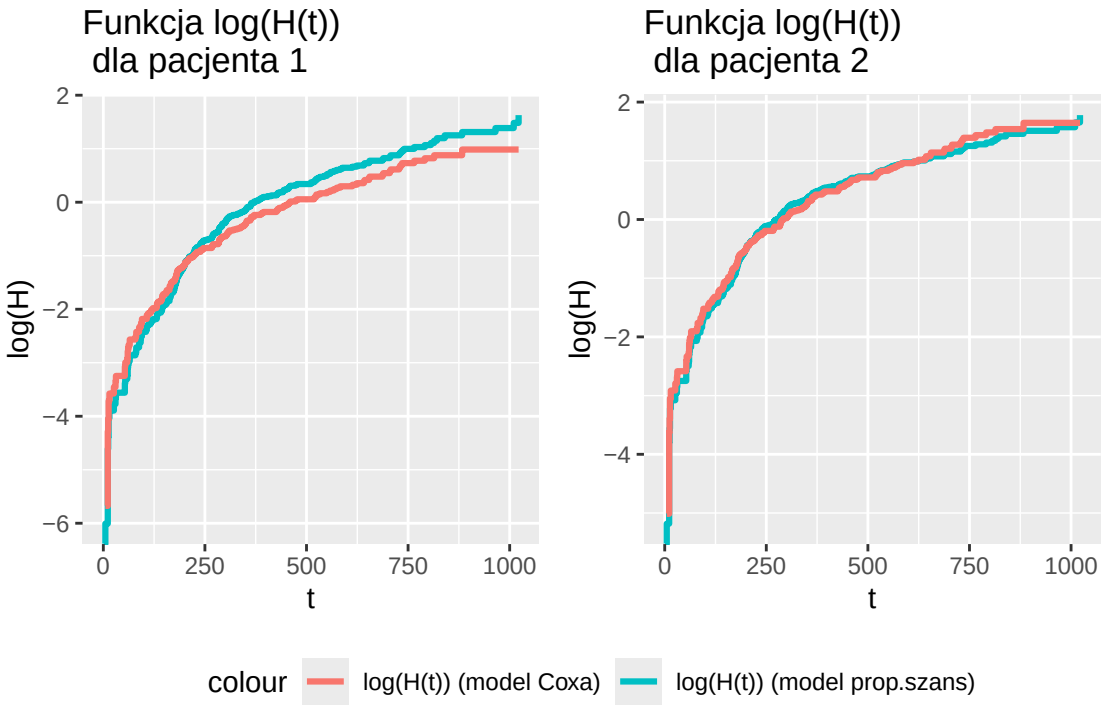
3.4 Zadanie 4



Rysunek 11: Porównanie oszacowania skumulowanej funkcji hazardu dla dwóch pacjentek

W przypadku pacjenta 1 widać, że skumulowany hazard przedstawiony na wykresie 11 (po lewejj stronie) estymowany w modelu proporcjonalnych szans rośnie wyraźnie szybciej niż w modelu proporcjonalnych hazardów Coxa. W konsekwencji model proporcjonalnych szans implikuje dla tego pacjenta mniej korzystną prognozę przeżycia, tj. większe narastające ryzyko zdarzenia w czasie, w porównaniu z modelem Coxa.

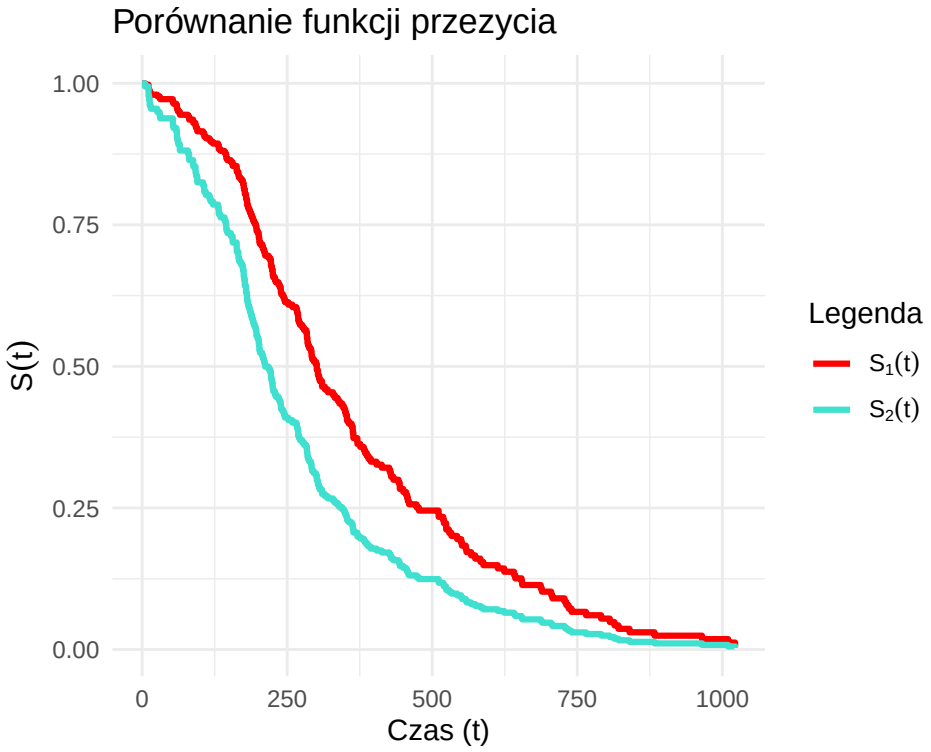
Dla pacjenta 2 krzywe (przedstawione po prawej stronie rysunku 11) skumulowanego hazardu są do około 625 dnia bardzo zbliżone, natomiast w dalszej części obserwacji zaczynają się rozchodzić. Różnice między predykcjami modeli narastają wraz z upływem czasu, jednak interpretację końcowego fragmentu należy traktować z ostrożnością — w ogonie rozkładu zwykle pozostaje niewiele obserwacji, a pojedyncze zdarzenia mogą powodować relatywnie duże, skokowe zmiany estymowanych krzywych.



Rysunek 12: Porównanie logarytmów z skumulowanej funkcji hazardu dla dwóch pacjentek

Analogiczne wnioski możemy wysnuć także na podstawie wykresu 12 logarytmów tych funkcji, ponieważ logarytm jest funkcją ściśle rosnącą i dla dodatnich wartości zachowuje relacje.

3.5 Zadanie 5



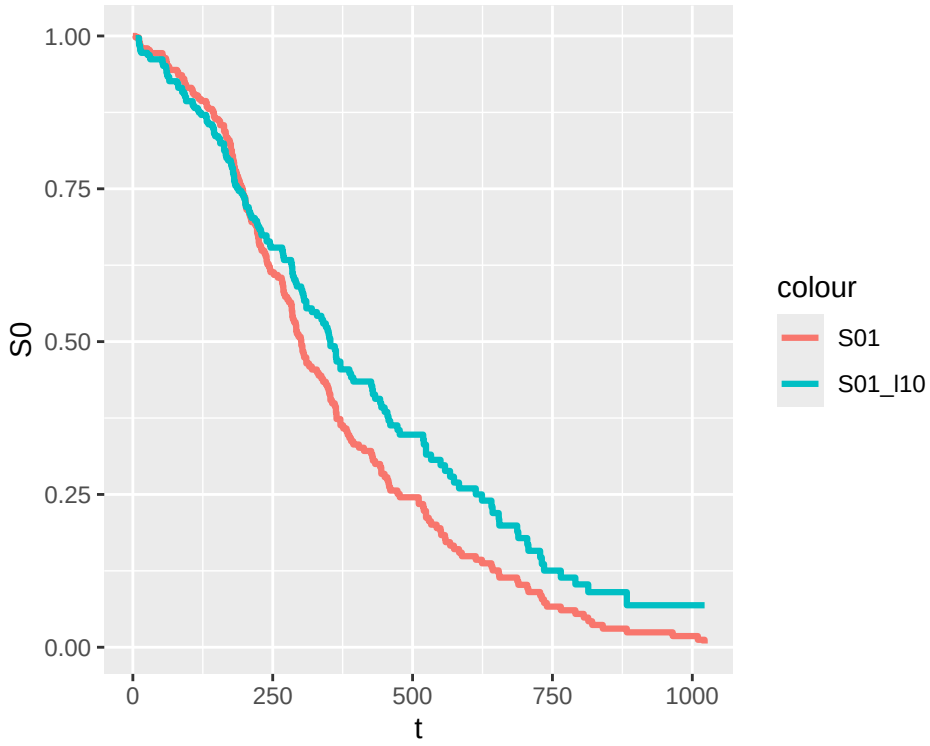
Rysunek 13: wykresy funkcji przeżycia dla pacjentów

Z wykresu 13 widać, że funkcja przeżycia dla pacjenta 2 maleje znacznie szybciej, co jest zgodne z charakterystyką danych.

Prawdopodobieństwo tego, że czas życia pacjenta 1 będzie większy od 300 dni wynosi 0.5032. Na liście 10 otrzymane prawdopodobieństwo wynosiło 0.5901. Prawdopodobieństwo dla modelu proporcjonalnych hazardów Coxa jest zatem większe niż dla modelu proporcjonalnych szans.

Prawdopodobieństwo tego, że czas życia pacjenta 2 będzie większy od 300 dni wynosi 0.3068 - model proporcjonalnych szans oraz 0.3598 - model Coxa. Prawdopodobieństwo dla modelu proporcjonalnych hazardów Coxa jest zatem większe niż dla modelu proporcjonalnych szans.

3.6 Zadanie 6



Rysunek 14: Porównanie wykresów oszacowanej funkcji przeżycia w modelu proporcjonalnych szans oraz modelu Coxa

Z wykresu 14 wynika, że estymowana funkcja przeżycia w modelu proporcjonalnych hazardów Coxa przyjmuje wyższe wartości niż w rozważanym modelu proporcjonalnych szans. Oznacza to, że model Coxa przewiduje korzystniejsze rokowanie (większe prawdopodobieństwo przeżycia od czasu t) w porównaniu z modelem proporcjonalnych szans.

4 Lista 12

4.1 Zadanie 1

W tym zadaniu omawiamy model **przyspieszonego czasu awarii**

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

4.1.1 a)

Dla zmiennej *age* w teście Walda otrzymujemy $p_{value} = 0.2052959 > 0.05$, co oznacza, że przy poziomie istotności $\alpha = 0.05$ nie ma podstaw do odrzucenia hipotezy $H_0 : \beta_{age} = 0$, a więc brak jest dowodów na istotność tej zmiennej w modelu.

Dla testu Ilorazu Wiarygodności otrzymaliśmy p_{value} =0.2013662, również > 0.05, co wskazuje, że uwzględnienie zmiennej *age* nie powoduje istotnej zmiany w dopasowanym modelu.

4.1.2 b)

Dla zmiennej *sex* w teście Walda otrzymujemy $p_{value} = 9.0513757 \times 10^{-4} < 0.05$, co oznacza, że przy poziomie istotności $\alpha = 0.05$ odrzucamy hipotezę $H_0 : \beta_{sex} = 0$. Wskazuje to, że zmienna *sex* ma istotny wpływ na czas do zdarzenia w modelu AFT, tzn. istotnie przyspiesza lub spowalnia czas awarii względem poziomu bazowego.

Dla testu Ilorazu Wiarygodności otrzymaliśmy $p_{value} = 5.9586684 \times 10^{-4} < 0.05$, co wskazuje, że uwzględnienie zmiennej *sex* ma istotny wpływ na dopasowany model.

4.1.3 c)

W teście ilorazu wiarygodności otrzymano p_{value} =0.0021387 < 0.05, więc przy poziomie istotności $\alpha = 0.05$ odrzucamy hipotezę zerową $H_0 : \beta_{ph.ecog=1} = \beta_{ph.ecog=2} = \beta_{ph.ecog=3} = 0$

(tj. brak wpływu poziomu ECOG na czas do zdarzenia względem poziomu bazowego, np. ECOG = 0). Oznacza to, że zmienna *ph.ecog* jest istotna w przyjętym modelu AFT, a poziom sprawności (ECOG) istotnie różnicuje czas przeżycia/czas awarii pacjentów.

4.2 Zadanie 2

```
model_podst <- coxph(Surv(time, status)~age +
                       as.factor(sex) +
                       as.factor(ph.ecog) +
                       ph.karno,
                       data = dane,)
```

4.2.1 a)

Dla zmiennej *age* w teście Walda otrzymujemy $p_{value} = 0.1839029 > 0.05$, co oznacza, że przy poziomie istotności $\alpha = 0.05$ nie ma podstaw do odrzucenia hipotezy $H_0 : \beta_{age} = 0$, a więc brak jest dowodów na istotność tej zmiennej w modelu.

Dla testu Ilorazu Wiarygodności otrzymaliśmy p_{value} =0.1803751, również > 0.05, co wskazuje, że uwzględnienie zmiennej *age* nie powoduje istotnej zmiany dopasowanego modelu.

4.2.2 b)

Dla zmiennej *sex* w teście Walda otrzymujemy $p_{value} = 8.609901 \times 10^{-4} < 0.05$, co oznacza, że przy poziomie istotności $\alpha = 0.05$ odrzucamy hipotezę $H_0 : \beta_{sex} = 0$. Wskazuje to, że zmienna *sex* ma istotny wpływ na czas do zdarzenia w modelu AFT, tzn. istotnie przyspiesza lub spowalnia czas awarii względem poziomu bazowego.

Dla testu Ilorazu Wiarygodności otrzymaliśmy $p_{value} = 1 < 0.05$, co wskazuje, że uwzględnienie zmiennej *sex* istotnie poprawia dopasowanie modelu w porównaniu z modelem bez tej zmiennej. Zatem *sex* jest istotnym predyktorem w modelu przyspieszonego czasu awarii, wpływając na skalę czasu przeżycia względem poziomu bazowego.

4.2.3 c)

W teście ilorazu wiarygodności otrzymano p_{value} =0.0021387 < 0.05, więc przy poziomie istotności $\alpha = 0.05$ odrzucamy hipotezę zerową $H_0 : \beta_{ph.ecog=1} = \beta_{ph.ecog=2} = \beta_{ph.ecog=3} = 0$

(tj. brak wpływu poziomu ECOG na czas do zdarzenia względem poziomu bazowego, np. ECOG = 0). Oznacza to, że zmienna *ph.ecog* jest istotna w przyjętym modelu proporcjonalnych hazardów Coxa, a poziom sprawności (ECOG) istotnie różnicuje czas przeżycia/czas awarii pacjentów.

4.3 Zadanie 3

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

4.3.1 a)

KROK 1

```
age <- anova(update(m, . ~ . - age), m)[["Pr(>Chi)"]][2]
sex <- anova(update(m, . ~ . - as.factor(sex)),      m)[["Pr(>Chi)"]][2]
ph.ecog <- anova(update(m, . ~ . - as.factor(ph.ecog)),  m)[["Pr(>Chi)"]][2]
ph.karno <- anova(update(m, . ~ . - ph.karno), m)[["Pr(>Chi)"]][2]

p_values <- data.frame(
  age = age,
  sex = sex,
  ph.ecog = ph.ecog,
  ph.karno = ph.karno
)

kable(p_values,digits = 3 ,caption = "Wartości p testu IW krok 1 AFT")
```

Tabela 6: Wartości p testu IW krok 1 AFT

age	sex	ph.ecog	ph.karno
0.201	0.001	0.002	0.133

Usuwamy zmienną *age*, ponieważ ma ona najwyższe p-value spośród rozważanych predyktorów ($p = 0.2013662 > 0.05$), co oznacza brak podstaw do uznania jej wpływu za istotny statystycznie w przyjętym modelu.

KROK 2

```
m <- survreg(Surv(time, status)~as.factor(sex) +
                  as.factor(ph.ecog) +
                  ph.karno,
                  data = dane,
```

```
                  dist = "weibull")

sex <- anova(update(m, . ~ . - as.factor(sex)),      m)[["Pr(>Chi)"]][2]
ph.ecog <- anova(update(m, . ~ . - as.factor(ph.ecog)),  m)[["Pr(>Chi)"]][2]
ph.karno <- anova(update(m, . ~ . - ph.karno), m)[["Pr(>Chi)"]][2]

p_values <- data.frame(
  sex = sex,
  ph.ecog = ph.ecog,
  ph.karno = ph.karno
)

kable(p_values,digits = 3 ,caption = "Wartości p testu IW kork 2 AFT")
```

Tabela 7: Wartości p testu IW kork 2 AFT

sex	ph.ecog	ph.karno
0.001	0.002	0.176

Usuwamy zmienną *ph.karno*, ponieważ ma ona najwyższe p-value spośród rozważanych predyktorów ($p = 0.1756437 > 0.05$), co oznacza brak podstaw do uznania jej wpływu za istotny statystycznie w przyjętym modelu.

KROK 3

Tabela 8: Wartości p testu IW krok3 AFT

sex	ph.ecog
0.001	0

Dla wszystkich pozostałych predyktorów otrzymano wartości $p < 0.05$, dlatego procedura selekcji (w oparciu o test IW) została zakończona i przyjęto model końcowy

4.3.2 b)

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

KROK 1

Tabela 9: Wartości AIC krok 1 AFT

wszystko	bez_age	bez_sex	bez_ph.ecoh	bez_ph.karno
2266.599	2266.231	2276.388	2275.252	2266.854

Ponieważ po usunięciu zmiennej *age* kryterium AIC przyjmuje najmniejszą wartość, uznajemy model bez *age* za lepiej dopasowany i dlatego *age* zostaje usunięta z modelu.

KROK 2

Tabela 10: Wartości AIC krok 2 AFT

wszystko	bez_sex	bez_ph.ecoh	bez_ph.karno
2266.231	2275.834	2274.984	2266.065

Ponieważ po usunięciu zmiennej *ph.karno* kryterium AIC osiąga najniższą wartość, wybieramy model bez *ph.karno* jako lepszy.

KROK 3

Tabela 11: Wartości AIC krok 3 AFT

wszystko	bez_sex	bez_ph.ecoh
2266.065	2274.735	2278.379

Usunięcie którejkolwiek z pozostałych zmiennych nie prowadzi do dalszego obniżenia wartości kryterium AIC, dlatego przyjmujemy uzyskany model jako model końcowy.

4.3.3 c)

Tutaj już korzystam z funkcji step, która autmatyzuje proces przedstawiony powyżej.

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

n <- sum(dane$status)

m_bic <- step(m, direction = "backward", k = log(n), trace = 0)

kable(m_bic$anova, caption = "Tabela kroków funkcji step kryterium BIC \n dla modelu AFT
```

Tabela 12: Tabela kroków funkcji step kryterium BIC dla modelu AFT

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

Według kryterium bayesowskiego (BIC) procedura selekcji prowadzi do takiego samego wniosku jak w poprzednich metodach: w modelu końcowym pozostają zmienne *sex* oraz *ph.ecog*. Eliminacja predyktorów przebiega analogicznie jak w przypadku selekcji opartej na AIC, przy czym decyzje podejmowane są na podstawie minimalizacji wartości BIC. Najpierw został odrzucony age a potem ph.karno.

4.4 Zadanie 4

```
m <- coxph(Surv(time, status)~age +
             as.factor(sex) +
             as.factor(ph.ecog) +
             ph.karno,
             data = dane)
```

4.4.1 a)

KROK 1

```
age <- anova(update(m, . ~ . - age), m)$"Pr(>|Chi|)"[2]
sex <- anova(update(m, . ~ . - as.factor(sex)),      m)$"Pr(>|Chi|)"[2]
ph.ecog <- anova(update(m, . ~ . - as.factor(ph.ecog)),  m)$"Pr(>|Chi|)"[2]
ph.karno <- anova(update(m, . ~ . - ph.karno), m)$"Pr(>|Chi|)"[2]

p_values <- data.frame(
  age = age,
  sex = sex,
  ph.ecog = ph.ecog,
  ph.karno = ph.karno
)

kable(p_values,digits = 3 ,caption = "Wartości p testu IW krok 1 coxph")
```

Tabela 13: Wartości p testu IW krok 1 coxph

age	sex	ph.ecog	ph.karno
0.18	0.001	0.004	0.189

Usuwamy zmienną *age*, ponieważ ma ona najwyższe p-value spośród rozważanych predyktorów ($p = 0.1803751 > 0.05$), co oznacza brak podstaw do uznania jej wpływu za istotny statystycznie w przyjętym modelu.

KROK 2

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

sex <- anova(update(m, . ~ . - as.factor(sex)),      m)$"Pr(>|Chi|)"[2]
ph.ecog <- anova(update(m, . ~ . - as.factor(ph.ecog)),  m)$"Pr(>|Chi|)"[2]
ph.karno <- anova(update(m, . ~ . - ph.karno), m)$"Pr(>|Chi|)"[2]

p_values <- data.frame(
  sex = sex,
  ph.ecog = ph.ecog,
  ph.karno = ph.karno
)

kable(p_values,digits = 3 ,caption = "Wartości p testu IW krok 2 coxph")
```

Tabela 14: Wartości p testu IW krok 2 coxph

sex	ph.ecog	ph.karno
0.001	0.003	0.246

Usuwamy zmienną *ph.karno*, ponieważ ma ona najwyższe p-value spośród rozważanych predyktorów ($p = 0.2455646 > 0.05$), co oznacza brak podstaw do uznania jej wpływu za istotny statystycznie w przyjętym modelu.

KROK 3

```
## [1] 0.001032973
```

Tabela 15: Wartości p testu IW krok 3 coxph

sex	ph.ecog
0.001	0

Dla wszystkich pozostałych predyktorów otrzymano wartości $p < 0.05$, dlatego procedura selekcji (w oparciu o test IW) została zakończona i przyjęto model końcowy

4.4.2 b)

```
m <- coxph(Surv(time, status)~age +
                  as.factor(sex) +
                  as.factor(ph.ecog) +
                  ph.karno,
                  data = dane)
```

KROK 1

Tabela 16: Wartości AIC krok 1 coxph

wszystko	bez_age	bez_sex	bez_ph.ecoh	bez_ph.karno
1458.537	1458.332	1468.226	1466.05	1458.266

Ponieważ po usunięciu zmiennej *age* kryterium AIC przyjmuje najmniejszą wartość, uznajemy model bez *age* za lepiej dopasowany i dlatego *age* zostaje usunięta z modelu.

KROK 2

Tabela 17: Wartości AIC krok 2 coxph

wszystko	bez_sex	bez_ph.ecoh	bez_ph.karno
1458.332	1467.841	1465.936	1457.68

Ponieważ po usunięciu zmiennej *ph.karno* kryterium AIC osiąga najniższą wartość, wybieramy model bez *ph.karno* jako lepszy.

KROK 3

Tabela 18: Wartości AIC krok 3 coxph

wszystko	bez_sex	bez_ph.ecoh
1457.68	1466.447	1470.137

Usunięcie którejkolwiek z pozostałych zmiennych nie prowadzi do dalszego obniżenia wartości kryterium AIC, dlatego przyjmujemy uzyskany model jako model końcowy.

4.4.3 c)

Tutaj już korzystam z funkcji step, która autmatyzuje proces przedstawiony powyżej.

```
m <- coxph(Surv(time, status)~age + as.factor(sex) + as.factor(ph.ecog) + ph.karno, dat

n <- sum(dane$status)

m_bic <- step(m, direction = "backward", k = log(n), trace = 0)

kable(m_bic$anova,
      caption = "Tabela kroków funkcji step kryterium BIC \n dla modelu propocjonalnych
```

Tabela 19: Tabela kroków funkcji step kryterium BIC dla modelu propocjonalnych hazardów Coxa

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

Według kryterium bayesowskiego (BIC) procedura selekcji prowadzi do takiego samego wniosku jak w poprzednich metodach: w modelu końcowym pozostają zmienne *sex* oraz *ph.ecog*. Eliminacja predyktorów przebiega analogicznie jak w przypadku selekcji opartej na AIC, przy czym decyzje podejmowane są na podstawie minimalizacji wartości BIC. Najpierw został odrzucony age a potem ph.karno.