Raport 3

Analiza przeżycia

Jakub Zdancewicz, Ksawery Józefowski

2026-01-31

Spis treści

1 Wstęp

Sprawozdanie jest podzielone na trzy części.

Część pierwsza poświęcona jest analizie danych pacjentów z zaawansowanym rakiem płuc, pochodzących ze zbioru `lung` dostępnego w pakiecie `survival`. W tej części skupiamy się na parametrycznych modelach regresji w analizie przeżycia, w szczególności na modelu przyspieszonego czasu awarii oraz modelu proporcjonalnych hazardów. Celem analizy jest oszacowanie parametrów rozważanych modeli, interpretacja uzyskanych współczynników oraz wyznaczenie oszacowań funkcji przeżycia i hazardu przy wybranych charakterystykach pacjentów.

Część druga skupia się na analizie tych samych danych, tym razem w kontekście semiparametrycznych modelów proporcjonalnych hazardów Coxa oraz proporcjonalnych szans. Celem analizy jest oszacowanie parametrów modelów, bazowej skumulowanej funkcji hazardu oraz bazowej funkcji przeżycia, a także interpretacja uzyskanych współczynników. Dodatkowo wyznaczymy oszacowania skumulowanej funkcji hazardu oraz funkcji przeżycia dla jednostek o wybranych charakterystykach.

Część trzecia zawiera natomiast weryfikację hipotez dotyczących istotności charakterystyk oraz równości rozkładów czasów życia dla różnych jednostek, przeprowadzoną przy użyciu testów Walda oraz testów ilorazu wiarogodności.

W części trzeciej

2 Parametryczne modele regresji w analizie przeżycia

Pierwszym krokiem analizy jest przygotowanie danych. W tym celu wczytujemy zbiór `lung` oraz wybieramy zmienne istotne z punktu widzenia dalszej analizy. Następnie usuwamy obserwacje brakujące.

```
dane <- lung[!(is.na(lung$ph.karno)) & !(is.na(lung$ph.ecog)),
             c("age", "sex", "ph.ecog", "ph.karno", "time", "status")]
sum(is.na(dane))
```

```
## [1] 0
```

Dodatkowo dokonujemy centrowania zmiennych ciągłych oraz zapisujemy ich średnie wartości. Celem tego zabiegu jest ułatwienie interpretacji parametrów modelu, tak aby ich wartości można było traktować jako odchylenie od średniej próbkowej.

```
age.mean <- mean(dane[, "age"])
ph.karno.mean <- mean(dane[, "ph.karno"])
dane["age"] <- dane["age"] - age.mean
dane["ph.karno"] <- dane["ph.karno"] - ph.karno.mean
```

Zakładamy, że funkcja przeżycia jednostki o charakterystyce zerowej S_0 ma rozkład Weibulla $\mathcal{W}](\lambda_0, \alpha_0)$. Funkcja ta przyjmuje postać

$$S_0(x) = \exp\Big(-\lambda_0 x^{\alpha_0}\Big)$$

2.1 Model przyspieszonego czasu awarii

Znając postać parametryczną funkcji przeżycia jednostki o charakterystyce zerowej, możemy zapisać postać funkcji przeżycia jednostki o charakterystyce z w modelu przyspieszonego czasu awarii:

$$S(x \mid z) = S_0\Big(\exp(\beta^\top z)\, x\Big)$$

Oszacujmy teraz parametry tego modelu, przyjmując za zmienną zależną zmienną `time`, a za

zmienne objaśniające zmienne `age`, `sex`, `ph.ecog` oraz `ph.karno`. Do estymacji parametrów wykorzystamy funkcję `survreg` z pakietu `survival`.

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

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)
```

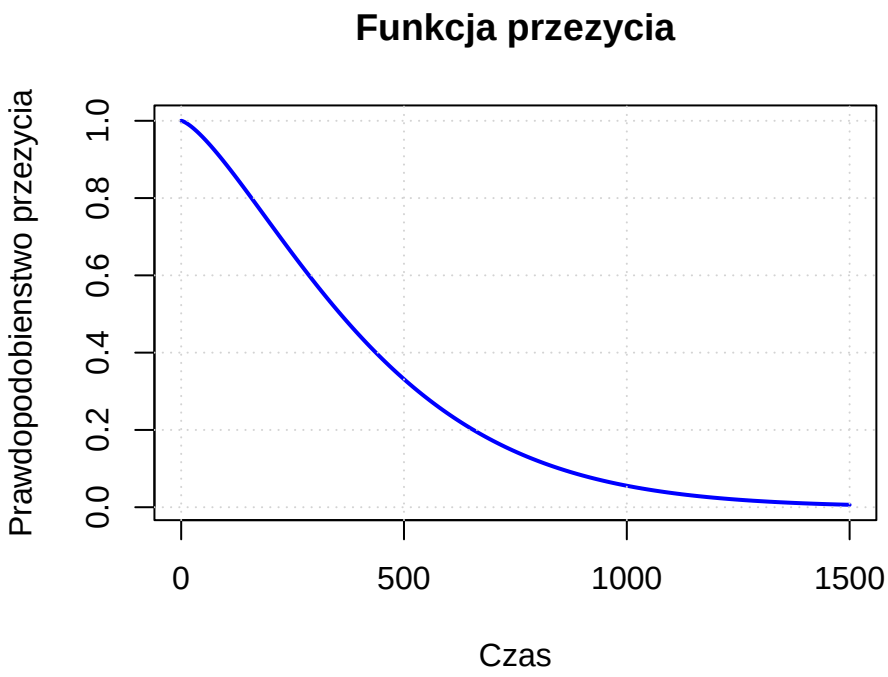
Współczynniki β modelu AFT mają następującą interpretację. Jeżeli $\exp\left(\beta^\top(z_1 - z_2)\right) > 1$ to jednostka o charakterystyce z_1 ma w każdej chwili większe prawdopodobieństwo wystąpienia zdarzenia przed czasem t niż jednostka o charakterystyce z_2. Odwrotnie, jeżeli $\exp\left(\beta^\top(z_1 - z_2)\right) < 1$ to prawdopodobieństwo to jest w każdej chwili mniejsze.

Następnym krokiem analizy jest wykorzystanie modelu z wyestymowanymi współczynnikami do oszacowania funkcji przeżycia jednostki o wybranych cechach: kobiety w wieku 70 lat, o charakterystyce ph.ecog $= 1$ oraz ph.karno $= 90$. Na podstawie uzyskanej funkcji przeżycia obliczymy prawdopodobieństwo, że czas życia tej kobiety przekroczy 300 dni.

```
z <- c(70-age.mean, 1, 1, 0, 0, 90-ph.karno.mean)
S_AFT <- function(x, z, beta, alpha, lambda) {
  exp(-lambda * exp(alpha * sum(beta[-1] * z)) * x^alpha)
}
S_AFT(300, z, beta_AFT, alpha_AFT, lambda_AFT)
```

```
## (Intercept)
##   0.5806085
```

Zatem szacowane prawdopodobieństwo, że kobieta o podanych charakterystykach przeżyje ponad 300 dni wynosi około 0.58.



Rysunek 1: Funkcja przeżycia odpowiadająca rozkładowi czasu życia kobiety w wieku 70 lat o zadanych charakterystykach w modelu przyspieszonego czasu awarii

Na rysunku 1 przedstawiamy funkcję przeżycia dla 70-letniej kobiety o charakterystyce ph.ecog $= 1$ oraz ph.karno $= 90$, oszacowaną na podstawie modelu przyspieszonego czasu awarii z wcześniej wyestymowanymi współczynnikami.

2.2 Model proporcjonalnych hazardów

Niech h_0 oznacza bazową funkcję hazardu odpowiadającą funkcji przeżycia S_0 jednostki o charakterystyce zerowej. Znając parametryczną postać bazowej funkcji hazardu

$$h_0(x) = \lambda_0 \alpha_0 x^{\alpha_0 - 1}$$

możemy zapisać funkcję hazardu jednostki o charakterystyce z w modelu proporcjonalnych hazardów w postaci:

$$h(x \mid z) = h_0(x) \, \exp\left(\beta^\top z\right).$$

Oszacujmy teraz parametry tego modelu, przyjmując za zmienną zależną zmienną `time`, a za zmienne objaśniające zmienne `age`, `sex`, `ph.ecog` oraz `ph.karno`. Do estymacji parametrów wykorzystamy funkcję `phreg` z pakietu `eha`.

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

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

Współczynniki β modelu PH mają następującą interpretację. Niech

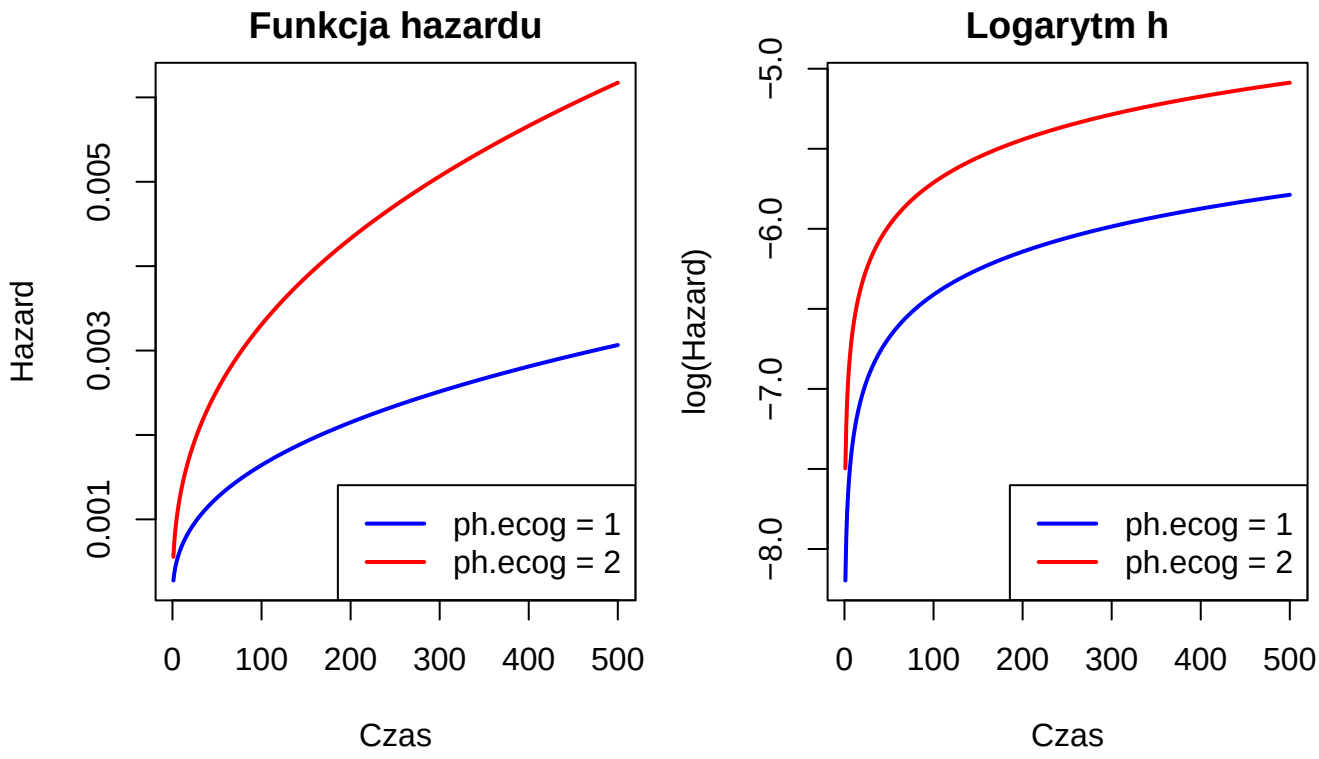
$$z_1 = (z_{11}, \ldots, z_{1k}, \ldots, z_{1p})^\top, \quad z_2 = (z_{21}, \ldots, z_{2k}, \ldots, z_{2p})^\top$$

będą dwoma wektorami charakterystyk takimi, że $z_{1k} = z_{2k} + 1$ oraz $z_{1i} = z_{2i}$ dla $i \neq k$. Wówczas zachodzi:

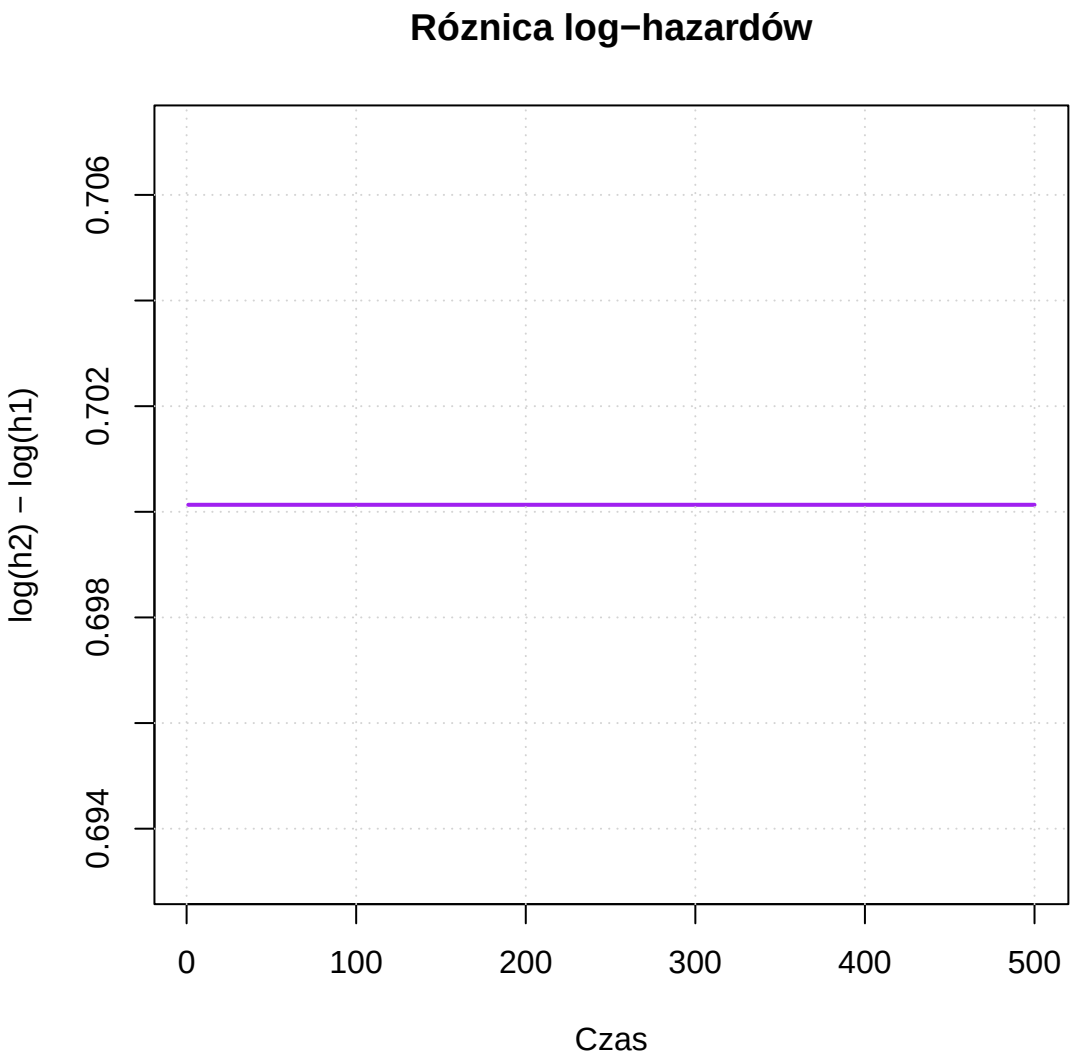
$$\frac{h(x \mid z_1)}{h(x \mid z_2)} = \exp(\beta_k).$$

Następnym krokiem analizy jest wykorzystanie modelu z wyestymowanymi współczynnikami do oszacowania funkcji hazardu jednostek o wybranych cechach: kobiety w wieku 70 lat o charakterystyce ph.ecog $= 1$ oraz ph.karno $= 90$, oraz kobiety w wieku 70 lat o charakterystyce ph.ecog $= 2$ oraz ph.karno $= 90$.

```r
z1 <- c(70-age.mean, 1, 1, 0, 0, 90-ph.karno.mean)
z2 <- c(70-age.mean, 1, 0, 1, 0, 90-ph.karno.mean)
h_PH <- function(x, z, beta, alpha, lambda) {
  lambda*alpha*x^(alpha-1) * exp(sum(beta * z))
}
```



Rysunek 2: Funkcje hazardu oraz ich logarytmy dla kobiet w wieku 70 lat o różnych wartościach zmiennej ph.ecog w modelu proporcjonalnych hazardów



Rysunek 3: Wykres różnic logarytmów funkcji hazardu kobiet w wieku 70 lat o określonych wartościach zmiennych ph.ecog i ph.karno w modelu proporcjonalnych hazardów

Na rysunku 2 przedstawiamy funkcje hazardu oraz ich logarytmy dla kobiet w wieku 70 lat o określonych charakterystykach. Z wykresów tych nie wynikają żadne wątpliwości co do przyjęcia modelu proporcjonalnych hazardów, ponieważ logarytmy hazardów wydają się być równoległe. Dodatkowo rysunek 3 jednoznacznie pokazuje, że różnica logarytmów hazardów jest stała w czasie, co potwierdza proporcjonalność hazardów.

Następnym krokiem będzie oszacowanie prawdopodobieństwa, że czas życia kobiet o podanych wcześniej charakterystykach przekroczy 300 dni. Do tego potrzebne są nam funkcje przeżycia jednostek o zadanych cechach. W modelu proporcjonalnych hazardów są one postaci

$$S(x \mid z) = \left[S_0(x)\right]^{\exp(\beta^\top z)}.$$

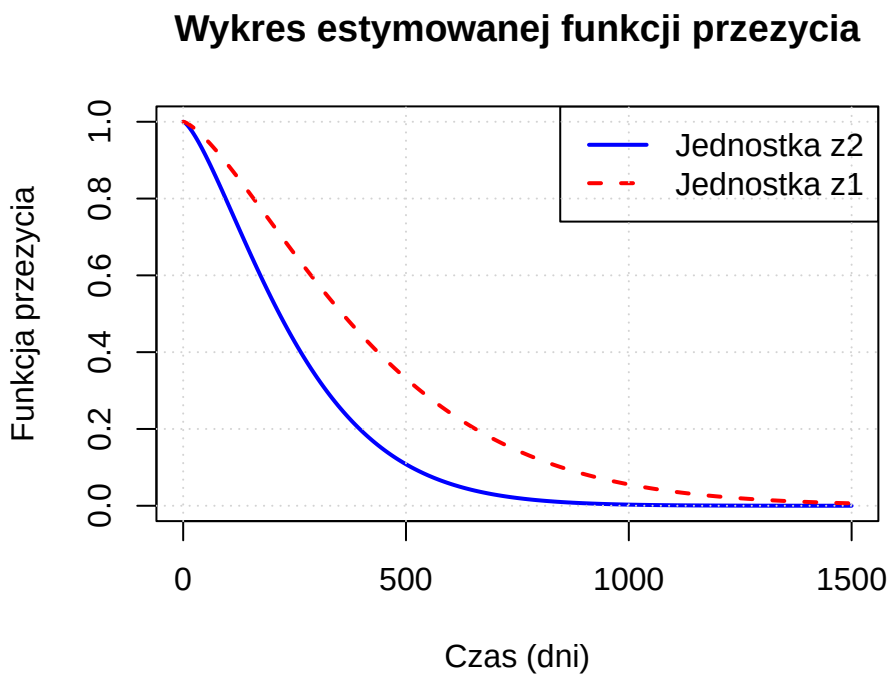
```
S_PH <- function(x, z, beta, alpha, lambda) {
  S0 <- exp(-lambda * x^alpha)
  S0^(exp(sum(beta * z)))
}

S_PH(300, z1, beta_PH, alpha_PH, lambda_PH)
```

```
## log(scale)
##  0.5806085
```

```
S_PH(300, z2, beta_PH, alpha_PH, lambda_PH)
```

```
## log(scale)
##  0.3345465
```

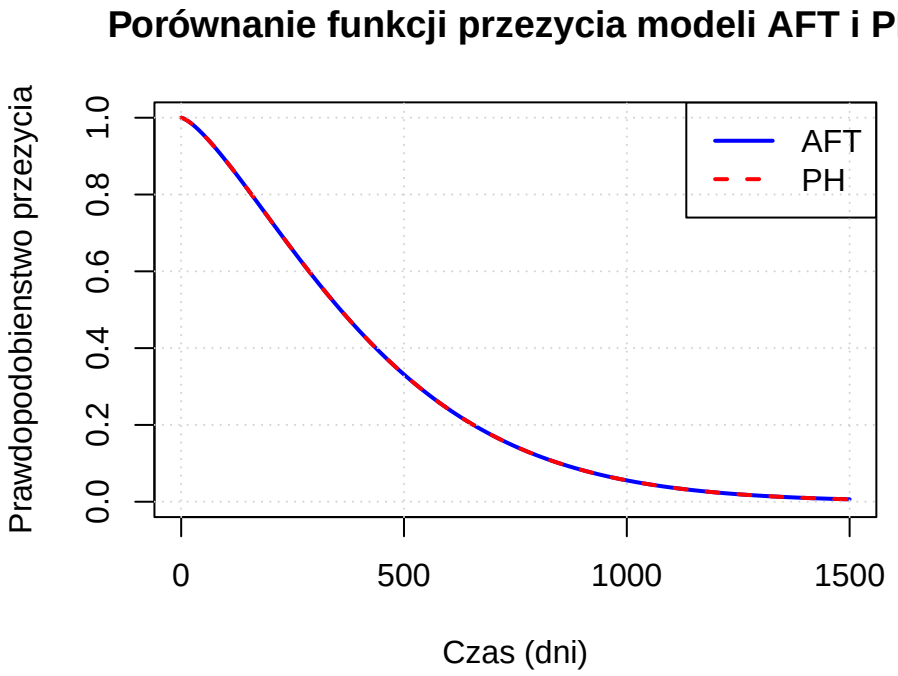


Rysunek 4: Wykres oszacowanych funkcji przeżycia obu kobiet o rozważanych charakterystykach w modelu PH

Na wykresie 4 przedstawiamy wykresy oszacowanych funkcji przeżycia.

Zauważmy, że prawdopodobieństwo przeżycia ponad 300 dni dla kobiety o charakterystyce ph.ecog = 1 oraz ph.karno = 90 jest takie samo w modelu przyspieszonego czasu awarii w

modelu proporcjonalnych hazardów.



Rysunek 5: Porównanie funkcji przeżycia kobiety w wieku 70 lat o charakterystyce ph.ecog = 1 oraz ph.karno = 90 obliczonych na podstawie modelu przyspieszonego czasu awarii (AFT) i modelu proporcjonalnych hazardów (PH)

Wykres 5 pokazuje, że wyestymowane funkcje przeżycia w obu modelach są identyczne dla tej kobiety.

3 Semiparametryczne modele regresji w analizie przeżycia

Do tej pory zajmowaliśmy się parametrycznymi modelami regresji, gdzie konieczne było przyjęcie postaci parametrycznej odpowiednich funkcji bazowych. Tym razem nie będziemy zakładać żadnego konkretnego bazowego rozkładu czasu życia i skorzystamy zarówno z metod parametrycznych, jak i nieparametrycznych do estymacji odpowiednio parametrów modelu oraz funkcji bazowego hazardu.

3.1 Model proporcjonalnych hazardów Coxa

Załóżmy, że h_0 jest bazową funkcją hazardu, o której niczego nie zakładamy, poza tym, że przyjmuje wartości nieujemne, zależy od czasu, ale nie zależy od charakterystyki z, oraz że

$$\int_0^\infty h_0(u)\,du = \infty.$$

Niech funkcja ψ zależy wyłącznie od wektora charakterystyk z, nie zależy od czasu t, przyjmuje wartości nieujemne i ma znaną postać parametryczną.

Wtedy, przyjmując semiparametryczny model proporcjonalnych hazardów Coxa, zakładamy, że rozkład czasu do wystąpienia zdarzenia dla jednostki o charakterystyce z ma funkcję hazardu postaci

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

Przyjmijmy również, że

$$\psi(z) = \exp(\beta^\top z).$$

Oszacujmy teraz parametry tego modelu, przyjmując za zmienną zależną zmienną `time`, a za zmienne objaśniające zmienne `age`, `sex`, `ph.ecog` oraz `ph.karno`. Do estymacji parametrów wykorzystamy funkcję `coxph` z pakietu `survival`.

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

Współczynniki β modelu Coxa mają następującą interpretację.
Niech charakterystyka z_1 ma postać

$$z_1 = (z_{11}, \ldots, z_{1k} + 1, \ldots, z_{1p})^\top,$$

a charakterystyka z_2 ma postać

$$z_2 = (z_{11}, \ldots, z_{1k}, \ldots, z_{1p})^\top.$$

Wówczas funkcje hazardu dla tych jednostek spełniają zależność:

$$h_{z_1}(t) = h_{z_2}(t)\,\exp(\beta_k).$$

Korzystając z relacji między funkcją hazardu a funkcją przeżycia, mamy:

$$S_{z_1}(t) = [S_{z_2}(t)]^{\exp(\beta_k)}.$$

Z tego wynika, że:

- jeśli $\beta_k > 0$, to $S_{z_1}(t) < S_{z_2}(t)$ dla każdego t,
- jeśli $\beta_k < 0$, to $S_{z_1}(t) > S_{z_2}(t)$ dla każdego t.

Oszacujmy teraz bazową skumulowaną funkcję hazardu oraz bazową funkcję przeżycia. W tym celu skorzystamy z funkcji `basehaz` dostępnej w pakiecie `survival` do wyznaczenia bazowej skumulowanej funkcji hazardu, a następnie wykorzystamy zależność

$$S_0(t) = \exp\Big(-H_0(t)\Big)$$

do obliczenia bazowej funkcji przeżycia.

```
model_Cox_H <- basehaz(model_Cox, centered=FALSE)

model_Cox_S0 <- exp(-model_Cox_H$hazard)
```

Następnym krokiem analizy jest wykorzystanie modelu z wyestymowanymi współczynnikami oraz oszacowaną bazową skumulowaną funkcją hazardu do wyznaczenia skumulowanej funkcji hazardu dla jednostek o wybranych cechach: kobiety w wieku 70 lat, o charakterystyce ph.ecog $= 1$ oraz ph.karno $= 90$, oraz kobiety w wieku 70 lat, o charakterystyce ph.ecog $= 2$ oraz ph.karno $= 90$.

W tym celu korzystamy z założonej postaci funkcji hazardu w modelu proporcjonalnych hazardów Coxa:

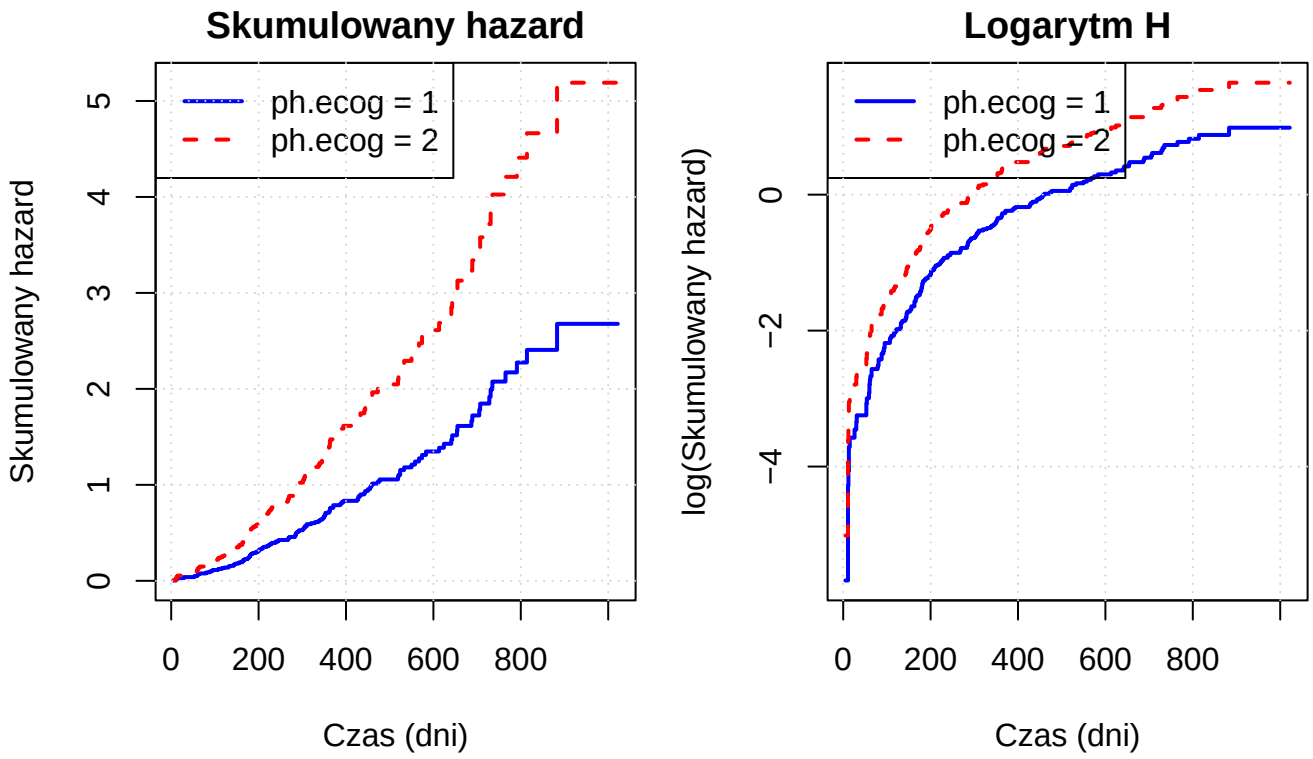
$$h(x \mid z) = h_0(x) \exp(\beta^\top z),$$

a po całkowaniu otrzymujemy skumulowaną funkcję hazardu:

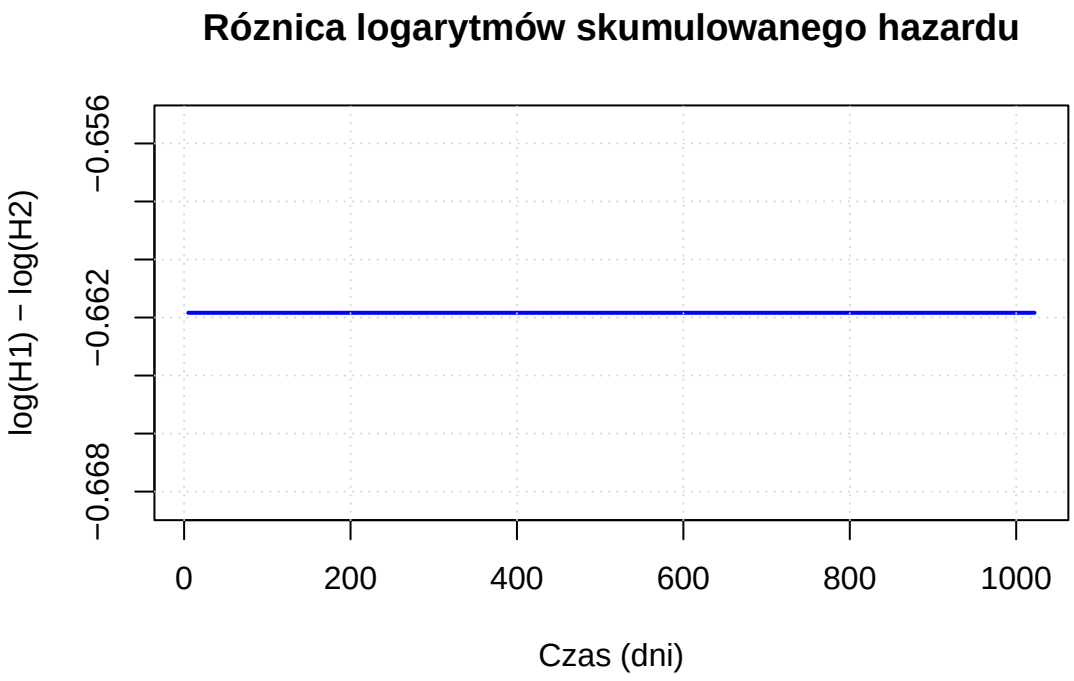
$$H(x \mid z) = H_0(x) \exp(\beta^\top z).$$

```
z1 <- c(70-age.mean, 1, 1, 0, 0, 90-ph.karno.mean)
z2 <- c(70-age.mean, 1, 0, 1, 0, 90-ph.karno.mean)

t_vec_cox <- model_Cox_H$time
H0 <- model_Cox_H$hazard
H1 <- H0 * exp(sum(beta_Cox * z1))
H2 <- H0 * exp(sum(beta_Cox * z2))
```



Rysunek 6: Skumulowana funkcja hazardu oraz jej logarytm dla kobiet w wieku 70 lat o różnych wartościach zmiennej ph.ecog w modelu proporcjonalnych hazardów Coxa



Rysunek 7: Różnica logarytmów skumulowanych funkcji hazardu dla kobiet w wieku 70 lat o różnych wartościach zmiennej ph.ecog w modelu proporcjonalnych hazardów Coxa

Na rysunku 6 przedstawiamy skumulowane funkcje hazardu oraz ich logarytmy dla kobiet w wieku 70 lat o określonych charakterystykach. Z wykresów tych nie wynikają żadne wątpliwości co do przyjęcia modelu proporcjonalnych hazardów, ponieważ logarytmy skumulowanych funkcji hazardów wydają się być równoległe. Dodatkowo wykres 7 jednoznacznie pokazuje, że różnica logarytmów hazardów jest stała w czasie, co potwierdza proporcjonalność skumulowanych funkcji hazardów hazardów.

Następnym krokiem będzie oszacowanie prawdopodobieństwa, że czas życia kobiet o podanych wcześniej charakterystykach przekroczy 300 dni. Do tego potrzebne są nam funkcje przeżycia jednostek o zadanych cechach. W modelu proporcjonalnych hazardów Coxa są one postaci

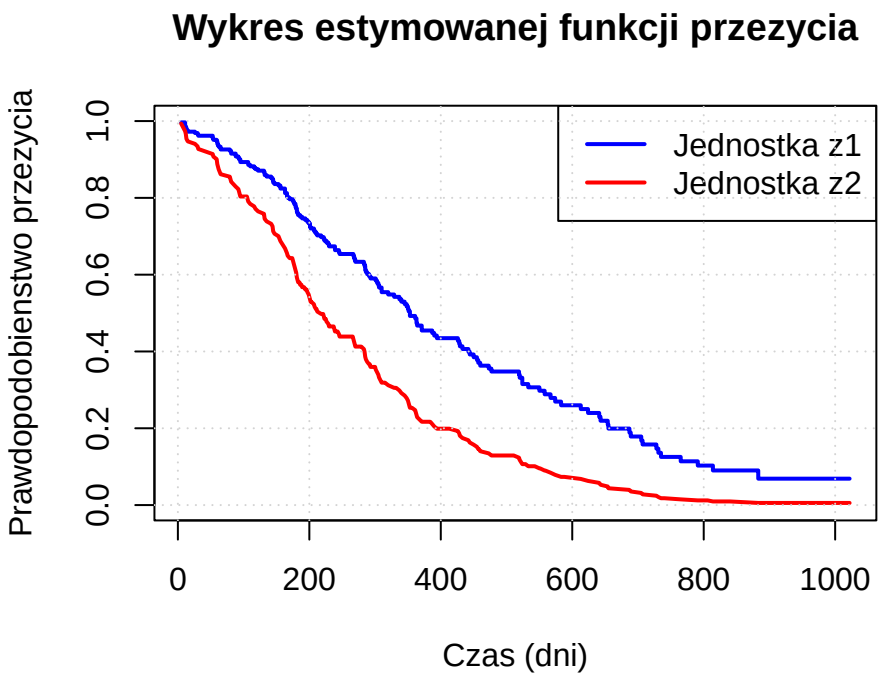
$$S(x \mid z) = \Big[S_0(x)\Big]^{\exp(\beta^\top z)}.$$

```
S_Cox <- function(z, beta) {
  model_Cox_S0^(exp(sum(beta * z)))
}
idx <- which.min(abs(t_vec_cox - 300))
S1 <- S_Cox(z1, beta_Cox)
S2 <- S_Cox(z2, beta_Cox)
S1[idx]
```

```
## [1] 0.5901344
```

```
S2[idx]
```

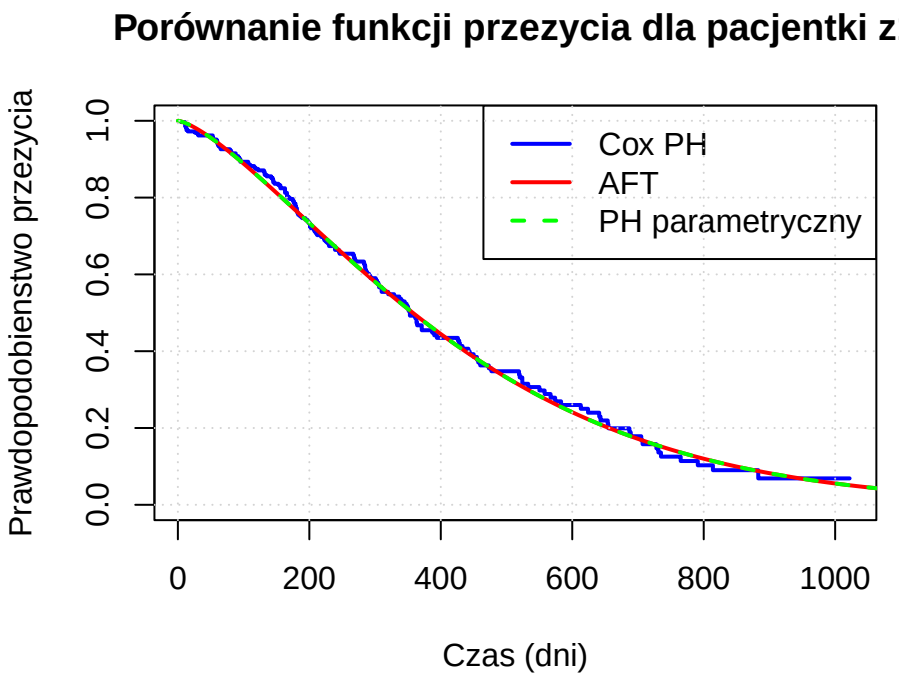
```
## [1] 0.3597682
```



Rysunek 8: Wykres oszacowanych funkcji przeżycia obu kobiet o rozważanych charakterystykach w modelu proporcjonalnych hazardów Coxa

Na rysunku8 przedstawiamy wykresy oszacowanych funkcji przeżycia.

Zauważmy, że prawdopodobieństwo przeżycia ponad 300 dni dla kobiety o charakterystyce ph.ecog = 1 oraz ph.karno = 90 jest wyższe w modelu proporcjonalnych hazardów Coxa niż we wcześniej rozważanych modelach parametrycznych.



Rysunek 9: Porównanie funkcji przeżycia dla pacjentki w wieku 70 lat o charakterystyce ph.ecog = 1 oraz ph.karno = 90 obliczonych na podstawie modelu proporcjonalnych hazardów Coxa, modelu AFT oraz modelu PH

Na wykresie 9 widać, że wyestymowane funkcje przeżycia, choć nieznacznie, różnią się od siebie.

3.2 Model proporcjonalnych szans

W tej części analizy rozpatrujemy model proporcjonalnych szans. W przeciwieństwie do modelu Coxa model PO opisuje zależność pomiędzy szansami, a nie pomiędzy hazardami. Jednocześnie jest to model semiparametryczny i, podobnie jak model Coxa, nie wymaga założenia konkretnej postaci rozkładu bazowego czasu przeżycia.

Semiparametrycznym modelem proporcjonalnych szans nazywamy model, w którym zakłada się, że szansa jednostki o charakterystyce z w chwili t ma postać

$$\theta_z(t) = \theta_0(t) \exp(\beta^\top z),$$

gdzie

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

a $S_0(t)$ oznacza bazową funkcję przeżycia.

W pierwszym kroku naszej analizy dostosowujemy zmienną cenzurującą do wymagań pakietu `timereg`, przekształcając wartości zmiennej `status` z $\{1, 2\}$ na $\{0, 1\}$:

```
dane$status <- dane$status - 1
```

Następnie oszacocujemy parametry modelu proporcjonalnych szans, przyjmując jako zmienną zależną zmienną `time`, natomiast jako zmienne objaśniające zmienne `age`, `sex`, `ph.ecog` oraz `ph.karno`.

```
model_PO <- prop.odds(Event(time, status) ~ age + as.factor(sex)
                      + as.factor(ph.ecog)
                      + ph.karno, data = dane, n.sim = 500, profile = 1)
beta_PO <- model_PO$gamma[, "estimate"]
```

Zauważmy, że przy założeniu modelu PO otrzymujemy prostą interpretację współczynników β. Mianowicie logarytm ilorazu szans dla dwóch jednostek o charakterystykach z_1 oraz z_2 ma postać

$$\log \frac{\theta_{z_1}(t)}{\theta_{z_2}(t)} = \beta^\top (z_1 - z_2),$$

co oznacza, że nie zależy on od czasu t.

Bazową funkcję przeżycia $S_0(t)$ wyznaczamy z wykorzystaniem funkcji `predict` z pakietu `timereg`, przyjmując zerowy wektor charakterystyk $z_0 = (0, 0, 0, 0, 0, 0)$.

```
s <- predict(model_PO, Z=c(0,0,0,0,0,0))
S0 <- s$S0[1, ]
```

Korzystając z zależności pomiędzy funkcją przeżycia a skumulowaną funkcją hazardu

$$H_0(t) = -\log\Big(S_0(t)\Big),$$

możemy wyznaczyć postać bazowej skumulowanej funkcji hazardu.

```
H0 <- -log(S0)
```

Dalszą analizę przeprowadzamy dla dwóch jednostek o następujących charakterystykach: kobieta w wieku 70 lat, o wartościach ph.ecog = 1 oraz ph.karno = 90, a także kobieta w wieku 70 lat, o wartościach ph.ecog = 2 oraz ph.karno = 90.

Funkcje przeżycia dla wybranych jednostek wyznaczamy na podstawie zależności

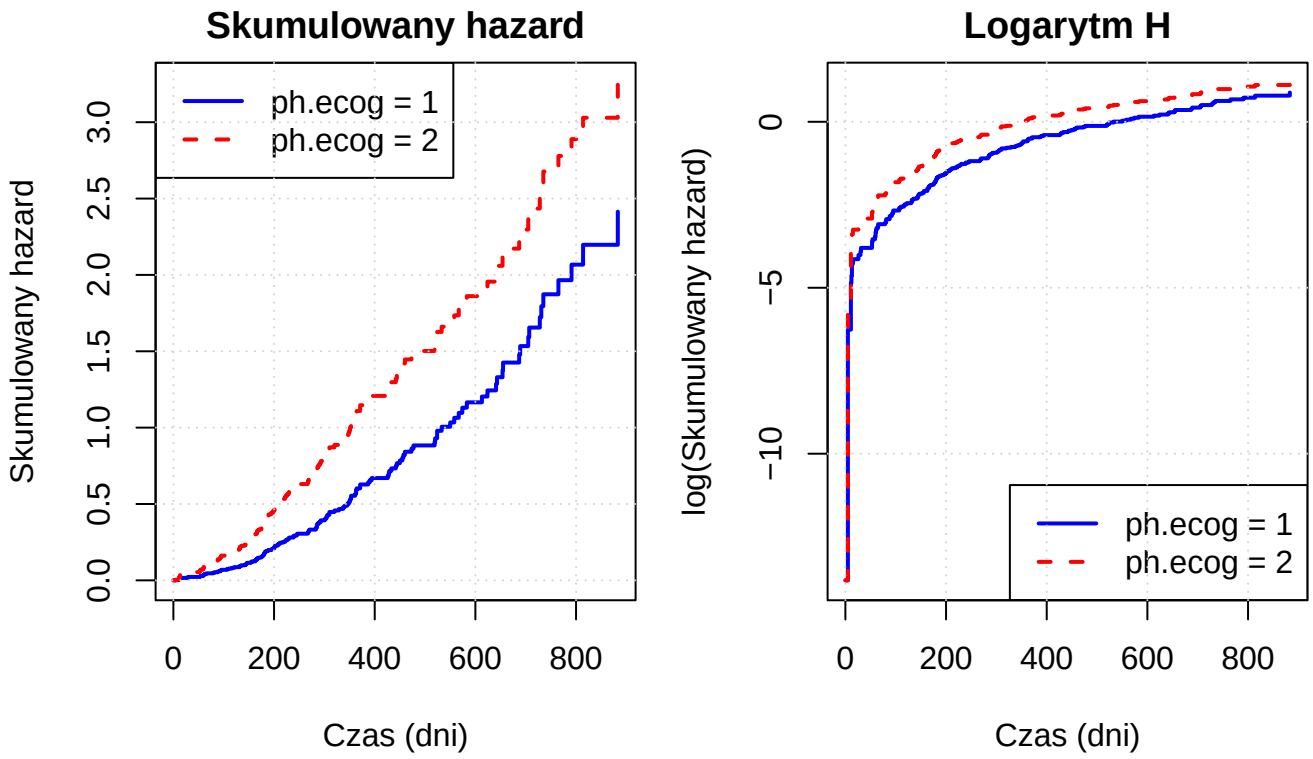
$$S_z(t) = \frac{1}{1 + \theta_0(t) \exp(\beta^\top z)}.$$

```
time_P0 <- s$time
theta0 <- (1 - S0) / S0

Sz1 <- 1 / (1 + theta0 * exp(sum(beta_P0 * z1)))
Sz2 <- 1 / (1 + theta0 * exp(sum(beta_P0 * z2)))
```

Skumulowaną funkcję hazardu dla jednostki o określonych cechach z wyznaczamy analogicznie jak w przypadku bazowej skumulowanej funkcji hazardu:

```
Hz1 <- -log(Sz1)
Hz2 <- -log(Sz2)
```

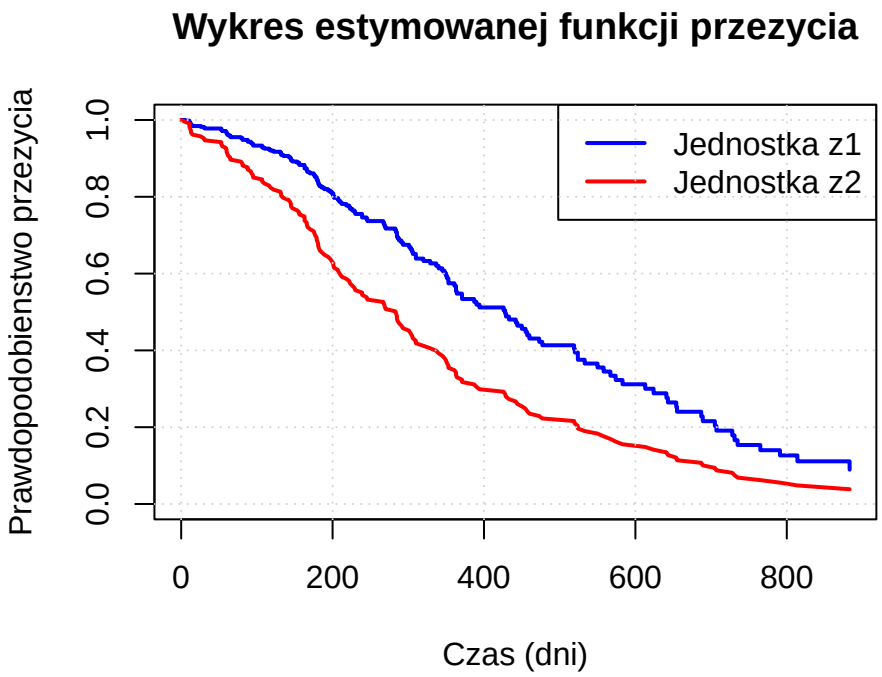


Rysunek 10: Skumulowana funkcja hazardu oraz logarytm skumulowanej funkcji hazardu dla kobiet w wieku 70 lat o różnych wartościach zmiennej extttph.ecog w modelu proporcjonalnych szans

Na rysunkach 6 oraz 10 przedstawiamy skumulowane funkcje hazardu oraz ich logarytmy dla kobiet w wieku 70 lat o określonych charakterystykach, obliczone odpowiednio w modelu proporcjonalnych hazardów Coxa oraz w modelu proporcjonalnych szans.

Zauważmy, że w przeciwieństwie do modelu proporcjonalnych hazardów Coxa, skumulowane funkcje hazardu w modelu PO nie wydają się być proporcjonalne i z czasem zbliżają się do siebie. Jest to zgodne z teoretycznymi własnościami modelu PO, w którym iloraz funkcji hazardu asymptotycznie dąży do 1. Ponadto widać, że wartości skumulowanej funkcji hazardu w modelu PO są mniejsze niż odpowiednie wartości w modelu PH Coxa dla obu kobiet o zadanych charakterystykach.

Kolejnym krokiem będzie wyznaczenie oszacowania funkcji przeżycia oraz wykorzystanie tego oszacowania do obliczenia prawdopodobieństwa przeżycia ponad 300 dni dla kobiet o podanych charakterystykach.



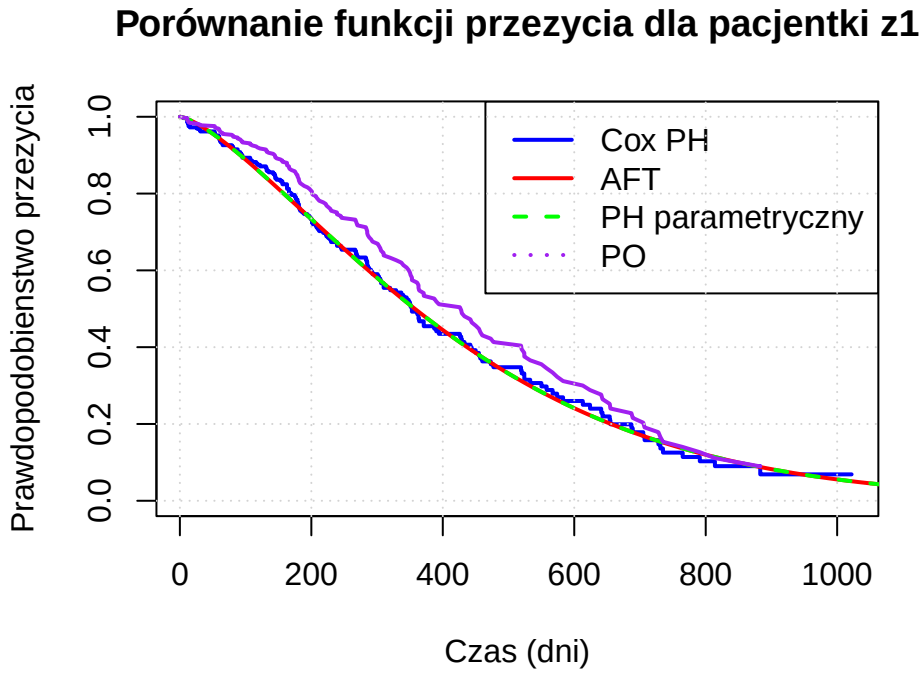
Rysunek 11: Wykres oszacowanych funkcji przeżycia obu kobiet o rozważanych charakterystykach w modelu propocjonalnych szans

Na rysunku 11 przedstawiamy wykresy oszacowanych funkcji przeżycia.

z1	z2
0.67	0.45

Tabela 1: Oszacowane prawdopodobieństwa przeżycia ponad 300 dni dla obu jednostek

W tabeli 1 przedstawiamy oszacowane prawdopodobieństwa przeżycia ponad 300 dni dla obu jednostek. Zauważamy, że w obu przypadkach prawdopodobieństwo przeżycia ponad 300 dni jest wyższe niż w modelu proporcjonalnych hazardów Coxa.



Rysunek 12: Porównanie funkcji przeżycia dla pacjentki w wieku 70 lat o charakterystyce ph.ecog = 1 oraz ph.karno = 90 obliczonych na podstawie modelu proporcjonalnych szans, modelu proporcjonalnych hazardów Coxa, modelu AFT oraz modelu PH

Na wykresie 12 widać, że wyestymowana funkcja przeżycia dla modelu proporcjonalnych szans znacząco różni się od pozostałych wyestymowanych funkcji przeżycia.

4 Testowanie hipotez dla parametrów modeli regresji

W tej części zajmiemy się analizą istotności wybranych cech pacjentów w badanym zbiorze danych w modelu przyspieszonego czasu awarii (AFT) z bazowym rozkładem Weibulla oraz modelu proporcjonalnych hazardów Coxa. Ponadto zweryfikujemy hipotezę o równości rozkładu czasu życia pacjentów o różnej sprawności ECOG, wykorzystując do tego testy Walda oraz testy ilorazu wiarogodności (IW).

Załóżmy poziom istotności $\alpha = 0.05$.

4.1 Testowanie hipotez dotyczących istotności zmiennych objaśniających w modelu

Na początku zweryfikujmy hipotezę, że zmienne `age` oraz `sex` nie są istotne w rozważanych modelach. W tym celu wykorzystamy testy Walda oraz testy ilorazu wiarogodności (IW).

```
# Wald
AFT_tab <- summary(model_AFT)[["table"]]
p_age_AFT <- AFT_tab["age", "p"]
p_sex_AFT <- AFT_tab["as.factor(sex)2", "p"]

Cox_tab <- summary(model_Cox)
p_age_Cox <- Cox_tab$coefficients["age", "Pr(>|z|)"]
p_sex_Cox <- Cox_tab$coefficients["as.factor(sex)2", "Pr(>|z|)"]

# IW
model_AFT_IW <- update(model_AFT, . ~ . - as.factor(sex))
anova_AFT_sex <- anova(model_AFT, model_AFT_IW)[["Pr(>Chi)"]][2]
model_AFT_IW <- update(model_AFT, . ~ . - age)
anova_AFT_age <- anova(model_AFT, model_AFT_IW)[["Pr(>Chi)"]][2]

model_Cox_IW <- update(model_Cox, . ~ . - as.factor(sex))
anova_Cox_sex <- anova(model_Cox, model_Cox_IW)[["Pr(>|Chi|)"]][2]
model_Cox_IW <- update(model_Cox, . ~ . - age)
anova_Cox_age <- anova(model_Cox, model_Cox_IW)[["Pr(>|Chi|)"]][2]
```

	age_Wald	sex_Wald	age_IW	sex_IW
AFT	0.2053	0.00091	0.2014	0.00060
Cox	0.1839	0.00086	0.1804	0.00063

Tabela 2: p-wartości testów Walda oraz testów ilorazu wiarogodności dla zmiennych `age` i `sex` w modelach AFT i Cox

W tabeli 2 przedstawiamy wyniki testów Walda oraz testów ilorazu wiarogodności (IW), zastosowanych do badania hipotezy zerowej H_0 o nieistotności zmiennych. Przy założonym poziomie istotności α wnioskujemy, że dla cechy `age` we wszystkich modelach i testach p-value $> \alpha$, co oznacza, że nie ma podstaw do odrzucenia hipotezy zerowej i wskazuje, że zmienna ta nie jest istotna.

Z drugiej strony dla cechy `sex` we wszystkich modelach i testach p-value $< \alpha$, co sugeruje odrzucenie hipotezy zerowej i uznanie tej cechy za istotną.

4.2 Badanie hipotezy o identycznych rozkładach czasu życia

W dalszej części zbadamy hipotezę H_0, że rozkłady czasu życia pacjentów o różnych poziomach sprawności `ECOG` $(0, 1, 2, 3)$ są takie same. Do tego celu wykorzystamy test ilorazu wiarogodności (IW), zaimplementowany w funkcji `anova`.

```
model_AFT_test <- update(model_AFT, . ~ . - as.factor(ph.ecog))
anova_AFT <- anova(model_AFT, model_AFT_test)[["Pr(>Chi)"]][2]

model_Cox_test <- update(model_Cox, . ~ . - as.factor(ph.ecog))
anova_Cox <- anova(model_Cox, model_Cox_test)[["Pr(>|Chi|)"]][2]
```

	AFT	Cox
p-value	0.0021	0.0036

Tabela 3: p-wartości dla modelu AFT i Cox weryfikujące równość rozkładów czasu życia pacjentów o różnych poziomach sprawności ECOG

Przedstawione w tabeli 3 p-wartości, przy przyjętym poziomie istotności, sugerują odrzucenie hipotezy H_0 dla obu modeli. Oznacza to, że rozkłady czasu życia różnią się dla pacjentów o różnych poziomach sprawności `ECOG` $(0, 1, 2$ i $3)$.

5 Wybór parametrów dla modeli regresji (Dodatkowe zadania)

W dalszej części analizy dokonamy wyboru zmiennych przy użyciu trzech metod:

- metody eliminacji,
- kryterium informacyjnego Akaike'a (AIC),
- kryterium Bayesa (BIC).

5.1 Model AFT

Najpierw wykorzystamy metodę eliminacji opartą na teście ilorazu wiarogodności (IW). Algorytm ten polega na rozpoczęciu analizy od modelu pełnego, zawierającego wszystkie rozważane zmienne objaśniające, a następnie sukcesywnym usuwaniu zmiennych, które nie są istotne statystycznie.

```
model_AFT_test <- update(model_AFT, . ~ . - as.factor(ph.ecog))
anova_AFT_1 <- anova(model_AFT, model_AFT_test)[["Pr(>Chi)"]][2]

model_AFT_test <- update(model_AFT, . ~ . - age)
anova_AFT_2 <- anova(model_AFT, model_AFT_test)[["Pr(>Chi)"]][2]

model_AFT_test <- update(model_AFT, . ~ . - age - as.factor(sex))
anova_AFT_3 <- anova(model_AFT, model_AFT_test)[["Pr(>Chi)"]][2]

model_AFT_test <- update(model_AFT, . ~ . - age - ph.karno)
anova_AFT_4 <- anova(model_AFT, model_AFT_test)[["Pr(>Chi)"]][2]
```

Tabela 4: p-values dla modelu AFT weryfikujące hipotezę o nie istotności parametru

	ph.ecog	age	sex	ph.karno
p-value	0.0021387	0.2013662	0.0013364	0.1767056

W pierwszej kolejności badamy pełny model z wyłączeniem zmiennej `ph.ecog`. Na podstawie testu na poziomie istotności $\alpha = 0.15$, którego wyniki przedstawiono w tabeli 4, stwierdzamy, że zmienna ta jest istotna.

Następnie rozważamy pełny model bez zmiennej `age`. Test zwraca p-value $> \alpha$, co sugeruje, że zmienna ta nie jest istotna, dlatego w kolejnych krokach zostaje usunięta z modelu.

Kolejny test przeprowadzamy dla poprzednio zmniejszonego modelu, nieuwzględniającego zmiennej `sex`. Wynik p-value wskazuje, że zmienna `sex` jest istotna, w związku z czym pozostaje w modelu.

Na koniec test wykonujemy dla modelu nieuwzględniającego zmiennych `ph.karno` oraz `age`. Wynik testu wskazuje na nieistotność zmiennej `ph.karno`, co pozwala ją wykluczyć z modelu.

Ostatecznie model AFT przyjmuje postać:

$$\text{Surv(time, status)} \sim \text{sex} + \text{ph.ecog}$$

Kolejnym etapem analizy jest wybór zmiennych w modelu AFT z wykorzystaniem kryteriów informacyjnych AIC oraz BIC. W tym celu zastosujemy funkcję `step`, realizującą selekcję krokową wsteczną.

```
AFT_AIC <- summary(step(model_AFT, direction = "backward",
                        k = 2))$table[, "p"]
AFT_BIC <- summary(step(model_AFT, direction = "backward",
                        k = log(nrow(dane))))$table[, "p"]
```

Tabela 5: p-wartości dla modelu AFT dla trzech metod wyboru zmiennych istotnych

	AIC	BIC
as.factor(sex)2	0.0017	0.0017
as.factor(ph.ecog)1	0.0410	0.0410
as.factor(ph.ecog)2	0.0000	0.0000
as.factor(ph.ecog)3	0.0549	0.0549

Punktem wyjścia dla obu eksperymentów był pełny model AFT, uwzględniający wszystkie rozważane zmienne objaśniające. W kolejnych krokach algorytm porównuje modele powstałe

poprzez usunięcie jednej ze zmiennych i wybiera wariant, dla którego wartość wybranego kryterium informacyjnego jest najmniejsza.

W przypadku kryterium AIC w pierwszym kroku procedura wskazała na usunięcie zmiennej `age`, ponieważ jej eliminacja prowadziła do poprawy dopasowania modelu. Następnie algorytm analizuje model bez zmiennej `age`, porównując go z modelami pozbawionymi kolejnych zmiennych. W kolejnym kroku algorytm eliminuje zmienną `ph.karno`, która nie poprawiała jakości modelu. Procedura zakończyła działanie w momencie, gdy dalsze usuwanie zmiennych prowadziłoby do wzrostu wartości kryterium AIC.

Analogiczne badanie przeprowadzono z wykorzystaniem kryterium BIC, które silniej karze za złożoność modelu. Również w tym przypadku, krok po kroku, algorytm najpierw usuwa zmienną `age`, a następnie zmienną `ph.karno`. Ostatecznie BIC wskazał ten sam zestaw zmiennych co AIC.

W tabeli 5 przedstawiono wartości p-value dla parametrów, które pozostały w modelach po zakończeniu procedury selekcji. Zarówno dla kryterium AIC, jak i BIC, zmienna `sex` oraz poszczególne poziomy zmiennej `ph.ecog` okazały się istotne statystycznie. Pozostawiamy w modelu również poziom `ph.ecog3`, mimo że jego wartość p-value była nieznacznie powyżej przyjętej granicy istotności, co uzasadniamy tym, że zmienna `ph.ecog` ma charakter zmiennej kategorycznej.

Ostateczny model AFT przyjął postać:

$$\text{Surv(time, status)} \sim \text{sex} + \text{ph.ecog}$$

5.2 Model Coxa

Analogicznie jak dla modelu AFT, w tabeli 6 przedstawiamy wartości p-value testu badającego nieistotność parametrów za pomocą metody eliminacji.

Kroki przeprowadzone w analizie są podobne do tych wykonanych w modelu AFT, w tym przypadku jako pierwsza eliminowana jest zmienna `ph.karno`, a następnie `age`. Uzyskany model końcowy składa się z identyczny zmiennych jak w przypadku modelu AFT.

```
model_Cox_test <- update(model_Cox, . ~ . - as.factor(ph.ecog))
anova_Cox_1 <- anova(model_Cox, model_Cox_test)[["Pr(>|Chi|)"]][2]

model_Cox_test <- update(model_Cox, . ~ . - ph.karno)
anova_Cox_2 <- anova(model_Cox, model_Cox_test)[["Pr(>|Chi|)"]][2]

model_Cox_test <- update(model_Cox, . ~ . - ph.karno - as.factor(sex))
anova_Cox_3 <- anova(model_Cox, model_Cox_test)[["Pr(>|Chi|)"]][2]

model_Cox_test <- update(model_Cox, . ~ . - ph.karno - age)
anova_Cox_4 <- anova(model_Cox, model_Cox_test)[["Pr(>|Chi|)"]][2]
```

Tabela 6: p-wartości dla modelu Coxa weryfikujące hipotezę o nie istotności parametru

	ph.ecog	ph.karno	sex	age
p-value	0.0036483	0.1885923	0.0019278	0.2077444

Analogiczną do modelu AFT procedurę selekcji zmiennych przy użyciu kryteriów informacyjnych AIC i BIC przeprowadzamy dla modelu proporcjonalnych hazardów Coxa, za pomocą funkcji `step`. Punkt wyjścia stanowił pełny model Coxa, uwzględniający zmienne `age`, `sex`, `ph.ecog` oraz `ph.karno`. Przebieg algorytmu był podobny jak w przypadku modelu AFT.

W kolejnych krokach algorytm eliminował zmienne, których usunięcie prowadziło do poprawy wartości odpowiedniego kryterium informacyjnego. Zarówno dla kryterium AIC, jak i BIC, w pierwszych etapach selekcji algorytm najpierw usuwa zmienną `ph.karno`, a następnie decyduje się na eliminację zmiennej `age`. Procedura zakończyła się w momencie, gdy dalsze upraszczanie modelu skutkowałoby wzrostem wartości kryterium AIC lub BIC.

W tabeli 7 przedstawiamy wartości p-value dla parametrów, które pozostały w modelu po zakończeniu procedury selekcji. Podobnie jak w analizie modelu AFT, zmienna `sex` oraz poszczególne poziomy zmiennej `ph.ecog` okazały się istotne statystycznie na przyjętym poziomie istotności, natomiast zmienne `age` i `ph.karno` zostały wyeliminowane z modelu.

Ostateczny model Coxa przyjmuje postać:

$$h(t \mid \text{sex}, \text{ph.ecog}) = h_0(t)\, \exp\Big(\beta_1 \cdot \text{sex} + \beta_2 \cdot \text{ph.ecog1} + \beta_3 \cdot \text{ph.ecog2} + \beta_4 \cdot \text{ph.ecog3}\Big),$$

```
Cox_AIC <- summary(step(model_Cox, direction = "backward",
              k = 2))$coefficients[, "Pr(>|z|)"]
Cox_BIC <- summary(step(model_Cox, direction = "backward",
              k = log(nrow(dane))))$coefficients[, "Pr(>|z|)"]
```

Tabela 7: p-wartości dla modelu Cox dla trzech metod wyboru zmiennych

	AIC	BIC
as.factor(sex)2	0.0014	0.0014
as.factor(ph.ecog)1	0.0358	0.0358
as.factor(ph.ecog)2	0.0000	0.0000
as.factor(ph.ecog)3	0.0440	0.0440