Analiza przeżycia
Raport 4
-
Modele proporcjonalnych hazardów Coxa

Romana Żmuda

5 lutego 2021

Spis treści

1 Zadanie do sprawozdania - Część 1 **3**
1.1 Zadanie 1 - Metoda proporcjonalnych hazardów Coxa 3
1.2 Zadanie 2 - Metoda proporcjonalnych hazardów Coxa z intepretacją 4

2 Zadanie do sprawozdania - Część 2 **6**
2.1 Zadanie 1 - Weryfikowanie istotności zmiennej *meal.cal* 6
2.2 Zadanie 2 - Weryfikowanie istotności zmiennej *pat.karno* 6
2.3 Zadanie 3 - Wybór zmiennych do modelu Coxa 7
2.3.1 Kryterium informacyjne AIC 7
2.3.2 Kryterium informacyjne BIC 9
2.4 Zadanie 4 - Wykres funkcji hazardu i przeżycia dla modelu z kryterium AIC . . 9
2.5 Zadanie 5 - Hipoteza o proporcjonalności hazardów 11

1 Zadanie do sprawozdania - Część 1

1.1 Zadanie 1 - Metoda proporcjonalnych hazardów Coxa

W tej części będziemy tworzyć modele semiparametryczne. które różnią się od modeli parametrycznych głównie tym, że zawierają w swojej postaci dwa elementy: jeden parametryczny, a drugi nieparametryczny. Są one modelami pośrednimi między modelami "czysto" parametrycznymi, w których zakłada się szczególną postać rozkładu obserwowalnych zmiennych losowych, a modelami "czysto" nieparametrycznymi, w których nie przyjmuje się żadnych założeń dotyczących postaci rozkładu.

Tak jak w poprzednim raporcie analizie poddamy zbiór danych *lung*, który dotyczy pacjentów z zaawansowanym rakiem płuc. . Zbiór zawiera informacje o 228 pacjentach, których zbiór charakterystyk obejmuje 8 następujących zmiennych:

- *inst* kod instytucji
- *time* czas przeżycia
- *status* cenzura (1. cenzura, 2. śmierć)
- *age* wiek
- *sex* płeć (1. mężczyzna, 2. kobieta)
- *ph.ecog* skala sprawności wg. lekarza (0-sprawność prawidłowa, 5-zgon)
- *ph.karno* skala sprawności wg. lekarza (sprawność prawidłowa - 100. zgon - 0)
- *pat.karno* skala sprawnośći wg. pacjenta
- *meal.cal* kalorie na posiłek
- *wt.loss* utrata masy ciała w ciągu ostatnich 6 miesięcy

Do modelu musimy założyć, że obserwowalne zmienne losowe mają rozkłady o ciągłych i różniczkowalnych dystrybuantach, a niektóre zmienne muszą być zmiennymi factor, poniżej utworzony model wykorzystując funkcję *coxph* z pakietu *survival*:

```
> dane <- data.frame(lung)
> dane$status <- as.factor(dane$status)
> dane$ph.ecog <-as.factor(dane$ph.ecog)
> dane$ph.karno <- as.factor(dane$ph.karno)
> dane$sex <- as.factor(dane$sex)
> dane$pat.karno <- as.factor(dane$pat.karno)
> model1<-coxph(formula=Surv(time,status == 2) ~ age + sex + ph.ecog + ph.karno
+               + pat.karno + meal.cal + wt.loss , data = dane)
```

Poniżej odpowiednie wartości do tworzenia modelu:

```
> #model1$coeff
> #summary(model1)
```

Zmienna	Wartość dopasowanych charakterystyk
age	5.871e-03
sex = 2	-6.082e-01
ph.ecog = 1	6.397e-01
ph.ecog = 2	1.320e+00
ph.ecog = 3	2.554e+00
ph.karno = 60	1.028e+00
ph.karno = 70	1.003e+00
ph.karno = 80	1.172e+00
ph.karno = 90	1.314e+00
ph.karno = 100	1.458e+00
pat.karno = 40	-3.519e-01
pat.karno = 50	7.532e-01
pat.karno = 60	1.228e-01
pat.karno = 70	-1.740e-01
pat.karno = 80	-2.811e-01
pat.karno = 90	-6.893e-02
pat.karno = 100	-5.681e-01
meal.cal	-4.431e-05
wt.loss	-1.394e-02

1.2 Zadanie 2 - Metoda proporcjonalnych hazardów Coxa z intepretacją

Jak w tytule zadania stworzymy model ze zmiennych: *sex*, *ph.ecog*, *ph.karno*, *pat.karno*, *wt.loss*.

```
> model2<-coxph(formula=Surv(time,status == 2) ~  sex + ph.ecog + ph.karno
+                + pat.karno + wt.loss , data = dane)
> summary(model2)

Call:
coxph(formula = Surv(time, status == 2) ~ sex + ph.ecog + ph.karno +
    pat.karno + wt.loss, data = dane)

  n= 210, number of events= 148
   (18 observations deleted due to missingness)

                coef exp(coef)  se(coef)      z Pr(>|z|)
sex2         -0.666328  0.513591  0.184476 -3.612 0.000304 ***
ph.ecog1      0.519718  1.681553  0.301850  1.722 0.085110 .
ph.ecog2      1.116277  3.053463  0.468803  2.381 0.017260 *
ph.ecog3      2.632819 13.912939  1.155307  2.279 0.022673 *
ph.karno60    0.864977  2.374951  0.670182  1.291 0.196822
ph.karno70    1.047251  2.849807  0.633549  1.653 0.098332 .
ph.karno80    1.365861  3.919095  0.647187  2.110 0.034819 *
ph.karno90    1.206609  3.342133  0.664854  1.815 0.069547 .
ph.karno100   1.260148  3.525945  0.728086  1.731 0.083493 .
pat.karno40   0.030517  1.030987  1.496812  0.020 0.983734
pat.karno50   1.058281  2.881415  1.201632  0.881 0.378478
pat.karno60   0.130915  1.139871  1.035814  0.126 0.899424
```

```
pat.karno70  -0.144684  0.865296  1.056975 -0.137 0.891122
pat.karno80  -0.325051  0.722491  1.057123 -0.307 0.758473
pat.karno90  -0.334883  0.715422  1.062647 -0.315 0.752655
pat.karno100 -0.507518  0.601988  1.076386 -0.472 0.637283
wt.loss      -0.014248  0.985853  0.007329 -1.944 0.051896 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

             exp(coef) exp(-coef) lower .95 upper .95
sex2            0.5136    1.94707   0.35776    0.7373
ph.ecog1        1.6816    0.59469   0.93063    3.0384
ph.ecog2        3.0535    0.32750   1.21828    7.6531
ph.ecog3       13.9129    0.07188   1.44551  133.9115
ph.karno60      2.3750    0.42106   0.63855    8.8331
ph.karno70      2.8498    0.35090   0.82326    9.8649
ph.karno80      3.9191    0.25516   1.10230   13.9339
ph.karno90      3.3421    0.29921   0.90803   12.3012
ph.karno100     3.5259    0.28361   0.84631   14.6901
pat.karno40     1.0310    0.96994   0.05485   19.3795
pat.karno50     2.8814    0.34705   0.27338   30.3695
pat.karno60     1.1399    0.87729   0.14968    8.6804
pat.karno70     0.8653    1.15567   0.10901    6.8685
pat.karno80     0.7225    1.38410   0.09099    5.7366
pat.karno90     0.7154    1.39778   0.08913    5.7423
pat.karno100    0.6020    1.66116   0.07301    4.9637
wt.loss         0.9859    1.01435   0.97179    1.0001

Concordance= 0.668  (se = 0.026 )
Likelihood ratio test= 42.64  on 17 df,   p=5e-04
Wald test            = 44.89  on 17 df,   p=3e-04
Score (logrank) test = 49.06  on 17 df,   p=6e-05
```

Sprawdźmy teraz intepretację współczynników przy zmiennych objaśniających *sex*, *ph.ecog*. Jeśli zachoruje kobieta to jej szansa śmierci wynosi 0,513591, więc śmierć mężczyzny jest prawie dwukrotnie wyższa niż kobiety. Zmienną referencyjną dla ph.ecog jest wartość wynoszącą 0. Zakładając, że według skali sprawności wg lekarza ph.ecog, pacjent osiągnie ph.ecog = 1, to wtedy jego śmiertelność wzrasta o 1.681553, gdy już ph.ecog = 2 to śmiertelność wzrasta o 3.053463 - 1.681553, gdy osiągnie wartość 3 wtedy śmiertelność wynosi prawie 14 razy więcej.

2 Zadanie do sprawozdania - Część 2

2.1 Zadanie 1 - Weryfikowanie istotności zmiennej *meal.cal*

W tym zadaniu zbadamy, czy zmienna *meal.cal* jest nieisotna w modelu. W skrócie mamy zweryfikować:

- $p > 0.05$ to znaczy, że nie ma podstaw do odrzucenia hipotezy H_0, wtedy zmienna badana nie jest istotna w modelu
- $p < 0.05$ to znaczy, że zmienna jest istotna w modelu

Skoro już wiemy, na co musimy zwrócić uwagę to zbudujmy nasz model i sprawdźmy istotność zmiennej.

```
> dane <- data.frame(lung)
> dane$status <- as.factor(dane$status)
> dane$ph.ecog <-as.factor(dane$ph.ecog)
> dane$ph.karno <- as.factor(dane$ph.karno)
> dane$sex <- as.factor(dane$sex)
> dane$pat.karno <- as.factor(dane$pat.karno)
> model_meal<-coxph(formula=Surv(time,status == 2) ~ age + sex + ph.ecog + ph.karno
+                   + pat.karno + meal.cal + wt.loss , data = dane)
> #summary(model_meal)
```

Zmienna	wartość p
meal.cal	0.8771

Widzimy, że wartość $p > 0.05$, dlatego zmienna w podanym modelu jest nieistotna i można ją odrzucić.

2.2 Zadanie 2 - Weryfikowanie istotności zmiennej *pat.karno*

Zadanie jest podobne do powyżej podanej treści (patrz zadanie 1 część 2) z małą zmienną w postaci badanej zmiennej na *pat.karno*. Rozważymy również wykorzystanie funkcji *anova*.

```
> model_pat<-coxph(formula=Surv(time,status == 2) ~ age + sex + ph.ecog + ph.karno
+                   + pat.karno + meal.cal + wt.loss , data = dane)
> #summary(model_pat)
> #anova(model_pat)
```

Zmienna	stopnie swobody	wartość $Pr(>\|Chi\|)$
pat.karno	7	0.616093

Widzimy, że wartość p jest powyżej 0.05, a więc znowu zmienna nie jest istotna w modelu i można ją usunąć. Widzimy również 7 stopni swobody, co oznacza ilości parametrów kryjących się pod zmienną *pat.karno*, pamiętajmy że jest to zmienna typu *factor*. Sprawdźmy na koniec co dzieje się, gdy usuniemy zmienne *pat.karno* oraz *meal.call*, zrobimy to porównując model ze zmiennymi do modelu bez nich.

```
> dane123 <- na.omit(dane)
> model_all<-coxph(formula=Surv(time,status == 2) ~ age + sex + ph.ecog + ph.karno
+                  + pat.karno + meal.cal + wt.loss , data = dane123)
> model_bez<-coxph(formula=Surv(time,status == 2) ~ age + sex + ph.ecog + ph.karno
+                  + wt.loss , data = dane123)
> anova(model_all, model_bez)

Analysis of Deviance Table
 Cox model: response is  Surv(time, status == 2)
 Model 1: ~ age + sex + ph.ecog + ph.karno + pat.karno + meal.cal + wt.loss
 Model 2: ~ age + sex + ph.ecog + ph.karno + wt.loss
   loglik  Chisq Df P(>|Chi|)
1 -490.87
2 -493.26 4.7794  8    0.7809
```

Widzimy, że model ze zmiennymi odrzucamy, gdyż $p > 0/05$, mimo że posiadana on aż 8 parametrów więcej, to nie wnoszą one do modelu na tyle by uznać je za istotne.

2.3 Zadanie 3 - Wybór zmiennych do modelu Coxa

2.3.1 Kryterium informacyjne AIC

W tym zadaniu mamy dokonać wyboru najlepszego modelu liniowego logarytmu czasu korzystając z kryterium informacyjnego Akaike'a (AIC). Poniżej odpowiednie wartości (Coefficients), czyli zmienne i ich wartości budujące model:

```
> model_AIC <- coxph(Surv(time, status == 2) ~ sex + age + ph.ecog +
+                  ph.karno + pat.karno + meal.cal
+                 + wt.loss, data = dane123)
> step(model_AIC)

Start:  AIC=1019.75
Surv(time, status == 2) ~ sex + age + ph.ecog + ph.karno + pat.karno +
    meal.cal + wt.loss

            Df    AIC
- pat.karno  7 1010.5
- ph.karno   5 1014.7
- meal.cal   1 1017.8
- age        1 1018.0
<none>         1019.8
- wt.loss    1 1020.6
- ph.ecog    3 1021.1
- sex        1 1026.3

Step:  AIC=1010.47
Surv(time, status == 2) ~ sex + age + ph.ecog + ph.karno + meal.cal +
    wt.loss

           Df    AIC
```

```
- ph.karno  5 1006.8
- meal.cal  1 1008.5
- age       1 1009.0
<none>        1010.5
- wt.loss   1 1011.7
- ph.ecog   3 1013.7
- sex       1 1018.3

Step:  AIC=1006.8
Surv(time, status == 2) ~ sex + age + ph.ecog + meal.cal + wt.loss

           Df    AIC
- meal.cal  1 1004.8
- age       1 1005.0
<none>        1006.8
- wt.loss   1 1007.4
- sex       1 1012.4
- ph.ecog   3 1015.1

Step:  AIC=1004.81
Surv(time, status == 2) ~ sex + age + ph.ecog + wt.loss

          Df    AIC
- age      1 1003.1
<none>        1004.8
- wt.loss  1 1005.5
- sex      1 1010.5
- ph.ecog  3 1013.3

Step:  AIC=1003.07
Surv(time, status == 2) ~ sex + ph.ecog + wt.loss

          Df    AIC
<none>       1003.1
- wt.loss  1 1003.8
- sex      1 1009.0
- ph.ecog  3 1014.0
Call:
coxph(formula = Surv(time, status == 2) ~ sex + ph.ecog + wt.loss,
    data = dane123)

               coef exp(coef)  se(coef)      z        p
sex2     -0.546262  0.579110  0.199796 -2.734 0.006255
ph.ecog1  0.385355  1.470136  0.235724  1.635 0.102098
ph.ecog2  1.106652  3.024217  0.285246  3.880 0.000105
ph.ecog3  2.202191  9.044808  1.044267  2.109 0.034959
wt.loss  -0.012451  0.987626  0.007694 -1.618 0.105588
```

```
Likelihood ratio test=23.16  on 5 df, p=0.000314
n= 167, number of events= 120
```

Ostatecznie charakterystykami tworzącymi model będą: sex = 2, całe ph.ecog, wt.loss. Poniżej odpowiednie wartości tworzące model dla tych zmiennych.

Zmienna	Coefficients
$sex = 2$	-0.546262
$ph.ecog = 1$	-0.546262
$ph.ecog = 2$	0.385355
$ph.ecog = 3$	2.202191
$wt.loss$	-0.012451

2.3.2 Kryterium informacyjne BIC

Ten podpunkt jest analogiczny co A, jednak tym razem skorzystamy z bayesowskiego kryterium informacyjnego (BIC) oraz z funkcji step. Poniżej odpowiednie wartości, należy pamiętać że w funkcji step dla kryterium BIC wspisujemy liczbę czystych rekordów. Poniżej odpowiednie wartości (Coefficient):

```
> #dim(dane123)
> model_BIC <- coxph(Surv(time, status == 2) ~ sex + age + ph.ecog +
+                  ph.karno + pat.karno + meal.cal
+                 + wt.loss, data = dane123)
> #step(model_BIC, k = log(167))
```

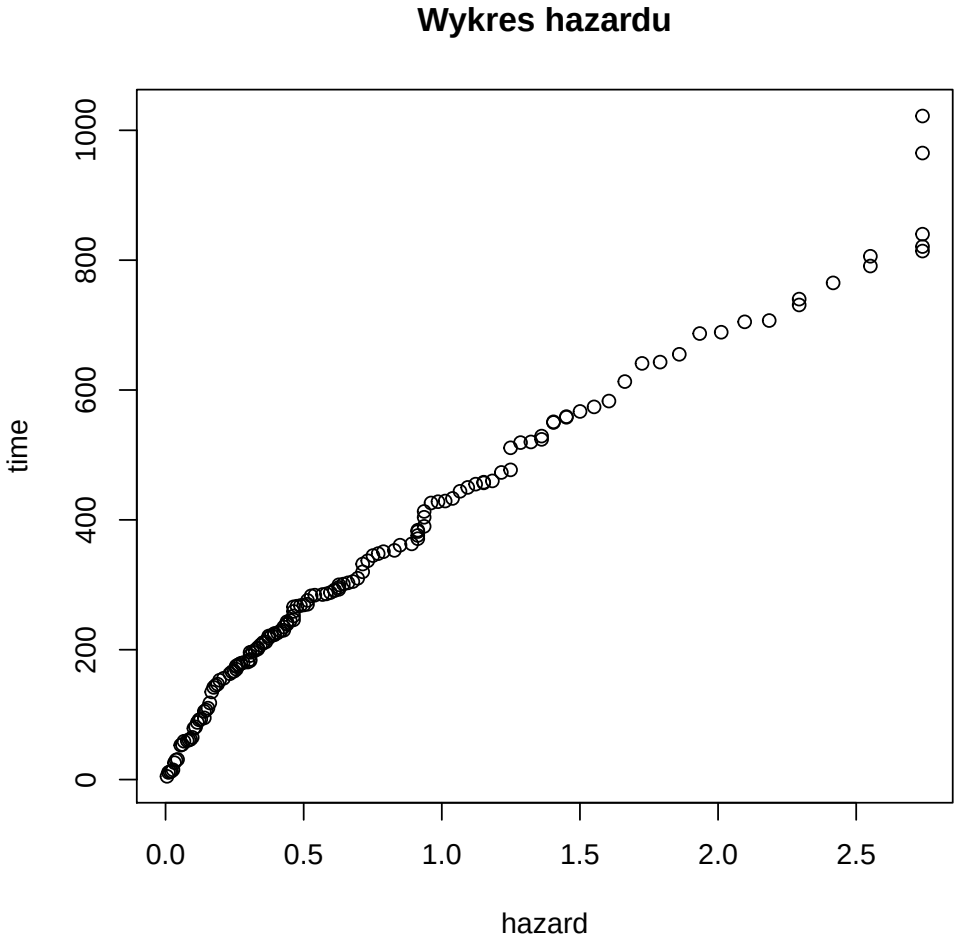
Ostatni krok pokazuje, że model zbudowany jest tylko sex = 2.

Zmienna	Coefficients
$sex = 2$	-0.4792

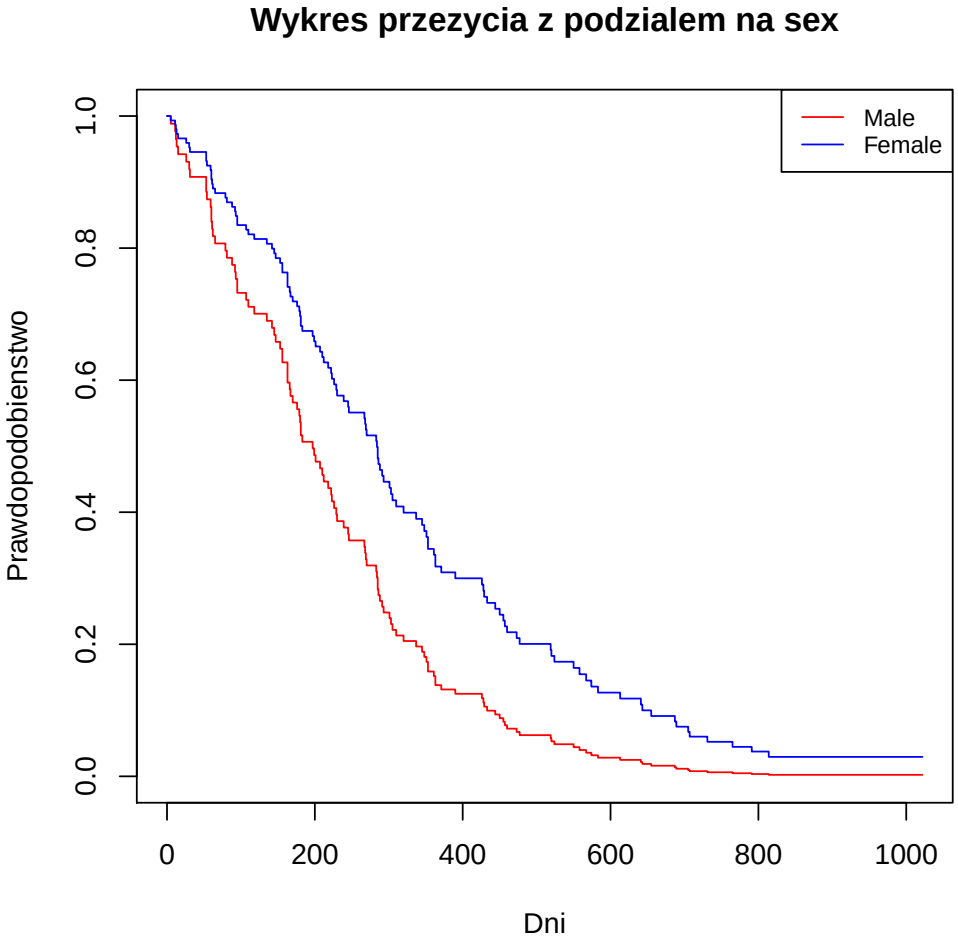
2.4 Zadanie 4 - Wykres funkcji hazardu i przeżycia dla modelu z kryterium AIC

W tym zadaniu mamy narysować wykres funkcji hazardu (zobacz rysunek 1) oraz funkcji przeżycia (zobacz rysunek 2) bazując na modelu z zadania 3 po dokonaniu wyboru zmiennych z kryterium AIC. Należy wspomnieć, że były to zmienne: *sex*, *ph.ecog*, *wt.loss*.

```
> model_AIC1 <- coxph(Surv(time, status == 2) ~ sex +
+                       ph.ecog + wt.loss, data = dane123)
> new <- data.table(sex = factor(c(1,2)), ph.ecog = factor(c(2,2)),
+                   wt.loss = c(15,15))
> fit <- survfit(model_AIC1, newdata = new)
```



Rysunek 1: Wykres funkcji hazardu



Rysunek 2: Wykres przeżycia z podziałem na zmienną *sex*

2.5 Zadanie 5 - Hipoteza o proporcjonalności hazardów

Aby móc powiedziec, czy spełnione jest założenie o proporcjonalnym hazardzie, używamy funkcji cox.zph. Funkcja ta : Testuje hipotezę zerową H_0 o spełnieniu założeń proporcjonalnego hazardu dla poszczególnych zmiennych objasniających. Będziemy testować model bez zmiennych pat.karno oraz meal.cal, gdyż w zadaniu pierwszym uznaliśmy, że model bez nich jest lepszy.

```
> dane <- data.frame(lung)
> dane$status <- as.factor(dane$status)
> dane$ph.ecog <-as.factor(dane$ph.ecog)
> dane$ph.karno <- as.factor(dane$ph.karno)
> dane$sex <- as.factor(dane$sex)
> model<-coxph(formula=Surv(time,status == 2) ~ age + sex + ph.ecog + ph.karno
+              + wt.loss , data = dane)
> hipoteza <- cox.zph(model)
> hipoteza
```

```
            chisq df    p
age      1.68e-04  1 0.99
```

```
sex      1.97e+00  1 0.16
ph.ecog  5.35e+00  3 0.15
ph.karno 5.50e+00  5 0.36
wt.loss  1.31e-01  1 0.72
GLOBAL   1.23e+01 11 0.34
```

Na podstawie uzyskanych danych, korzystając z testu Grambscha i Therneau'a, na poziomie istotności 0.05, nie ma podstaw do odrzucenia hipotezy o proporcjonalności hazardów w przyjętym przez nas modelu. Wartość poziomu krytycznego w tym teście wynosi 0.34 .
Funkcja cox.zph sprawdza założenie proporcjonalności, używając reszt Schoenfelda względem czasu przekształconego. Bardzo małe wartości p wskazują, że istnieją współczynniki zależne od czasu, którymi należy się zająć. Oznacza to, że założenie proporcjonalności nie sprawdza liniowości - model Cox PH jest półparametryczny, a zatem nie przyjmuje żadnych założeń co do formy zagrożenia. Założeniem proporcjonalności jest to, że współczynnik hazardu jednostki jest względnie stały w czasie i to właśnie testuje cox.zph.

Jeśli zmienna towarzysząca łamie założenie, może wymagać naprawy, ponieważ istnieją współczynniki zależne od czasu. W podanym modelu żadna zmienna nie jest poniżj przyjętego poziomu p.