

Praca domowa nr 2 - raport

Gabriel Bożek, Luiza Dobosz, Anna Sikorska, Aneta Skorupska

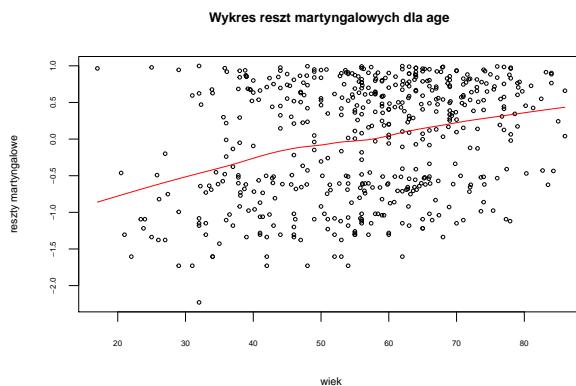
07.04.2020

Otrzymaliśmy dane dotyczące 517 chorych na chłoniaka leczonych radioterapią i, ewentualnie, chemioterapią. Dane zawierały następujące zmienne:

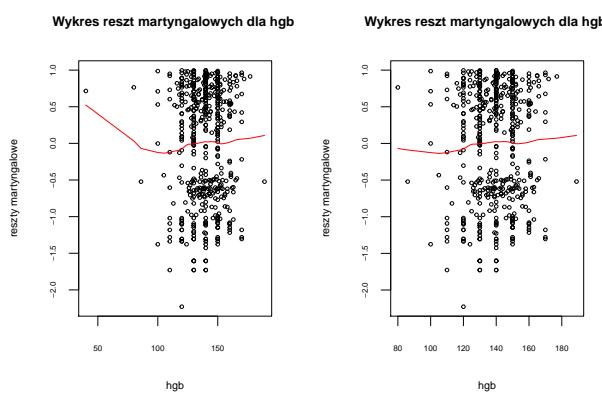
- stnum - identyfikator chorego
- age - wiek chorego w chwili diagnozy (lata)
- clinstg - stopień zaawansowania choroby (1 – I, 2 – II)
- hgb - poziom hemoglobiny (g/l)
- chemo - wskaźnik leczenia chemioterapią (0 – nie, 1 – tak)
- dftime - czas do wystąpienia nawrotu choroby lub zgonu (lata)
- dfstat - wskaźnik zdarzenia (0 – żył bez nawrotu, 1 – nawrót i/lub zgon)

Chcemy sprawdzić, które ze zmiennych charakteryzujących chorego mają wpływ na czas przeżycia bez nawrotu choroby, tzn. czas do wystąpienia nawrotu lub zgonu. Do analizy tych zmiennych użyjemy modelu proporcjonalnych hazardów. We wszystkich testach przyjmiemy poziom istotności równy 0.05.

Zauważmy, że mamy dwie zmienne ciągłe: age i hgb. Zastanówmy się jaką ich postać funkcjonalną przyjąć do modelu. W tym celu dokonamy analizy wykresów reszt martyngałowych.



Otrzymujemy w przybliżeniu zależność liniową, dlatego jako przekształcenie zmiennej age wybieramy identyczność.



Zauważmy, że na pierwszym wykresie na kształt krzywej ma wpływ obserwacja odstająca. Jeśli narysujemy wykres bez tej obserwacji to otrzymamy w przybliżeniu linię prostą. Nie należy jednak usuwać obserwacji ze zbioru danych, ponieważ nie musi być ona błędna. Po konsultacji z ekspertem z dziedziny medycyny dowiedzieliśmy się, że tak niski poziom hemoglobiny (40 g/l) jest możliwy przy chłoniaku. Sprawdźmy jednak, czy przy założeniu liniowej postaci funkcjonalnej obserwacja odstająca zmieni znacząco jej współczynnik kierunkowy. Po lewej stronie mamy wyniki dla pełnego zbioru danych, a po prawej stronie dla modelu z usuniętą obserwacją dla $\text{hgb}=40$.

```
Call:
coxph(formula = Surv(dftime, dfstat) ~ age + hgb + clinstg +
    chemo, data = df)

      coef exp(coef)   se(coef)      z     p
age    0.032114  1.032635 0.004491 7.151 8.64e-13
hgb    0.004925  1.004937 0.003829 1.286 0.198341
clinstg 0.476039  1.609686 0.125467 3.794 0.000148
chemo  -0.539191  0.583220 0.171691 -3.140 0.001687

Likelihood ratio test=69.26 on 4 df, p=3.247e-14
n= 517, number of events= 324

Call:
coxph(formula = Surv(dftime, dfstat) ~ age + hgb + clinstg +
    chemo, data = df2)

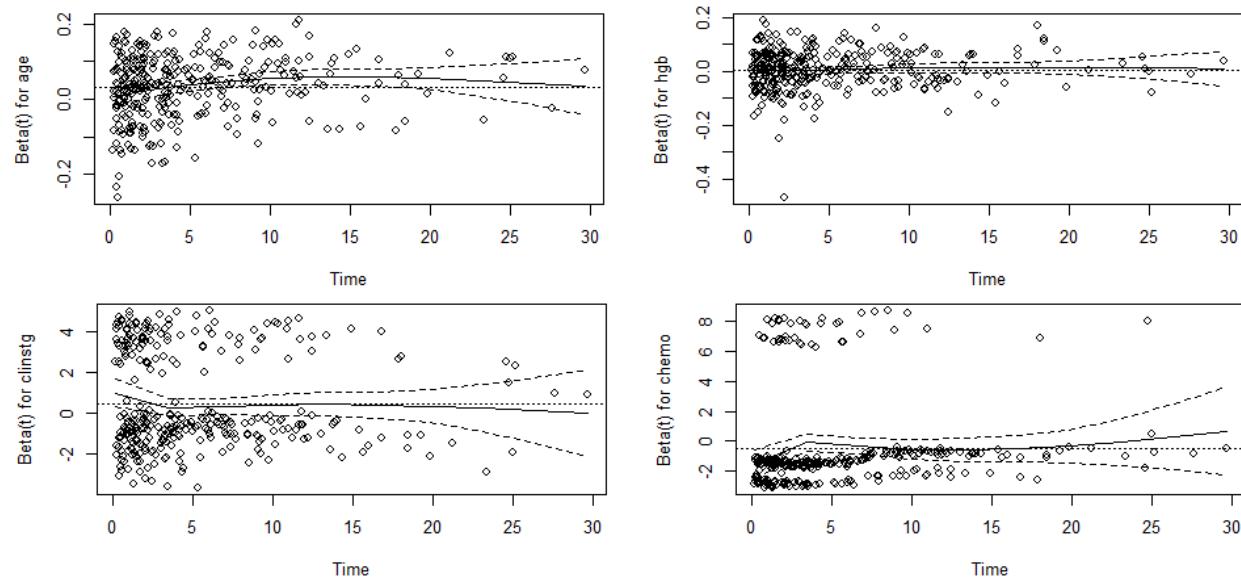
      coef exp(coef)   se(coef)      z     p
age    0.032201  1.032725 0.004505 7.148 8.8e-13
hgb    0.006118  1.006137 0.003897 1.570 0.116370
clinstg 0.483895  1.622381 0.125634 3.852 0.000117
chemo  -0.538763  0.583469 0.171776 -3.136 0.001710

Likelihood ratio test=69.73 on 4 df, p=2.585e-14
n= 516, number of events= 323
```

Sprawdziliśmy, że różnice pomiędzy współczynnikami mieszą się w ich błędach standardowych. W takim razie obserwacja nie wpływa znacząco na współczynniki modelu z liniową postacią funkcjonalną zmiennej hgb i przy tej postaci zostajemy.

Na podstawie wyników testu Walda nie jesteśmy w stanie powiedzieć, że hemoglobina ma istotny wpływ na czas przeżycia. Pozostałe zmienne mają istotny wpływ. Zwróćmy uwagę, że z reszt martyngałowych wynikało, że postać funkcyjna hemoglobiny mogłyby być nawet funkcją stałą.

Sprawdźmy teraz, czy założenia modelu PH są spełnione. W pierwszym kroku oceniliśmy wykresy krzywych przeżycia dla każdej zmiennej. Patrzyliśmy, czy wykresy się krzyżują oraz po transformacji log-log, czy są równoległe. Nie wszystkie wykresy wskazywały spełnienie założeń. Jako wiążące przyjmujemy wyniki testu Schoenfelda z transformacją identycznościową czasu wynikającą z postaci wykresów skalowanych reszt Schoenfelda.



```
##          chisq df      p
## age      10.95 1 0.00094
## hgb      1.25  1 0.26395
## clinstg  2.58  1 0.10857
## chemo    0.52  1 0.47081
## GLOBAL   13.29 4 0.00992
```

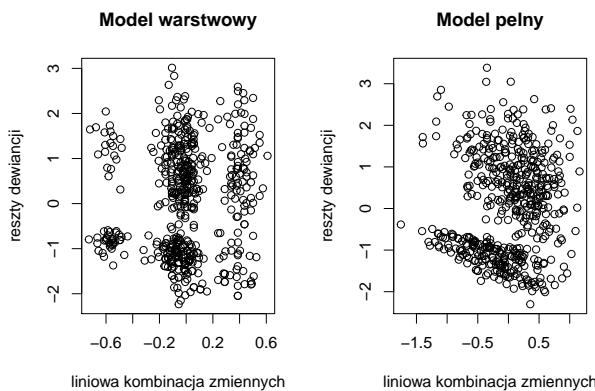
Wynik testu skłania nas do stwierdzenia, że wiek nie spełnia założeń modelu PH. Z wykresu skalowanych reszt Schoenfelda widzimy, że zależność współczynnika od czasu nie jest znacząca. Stworzymy więc model z warstwami wieku i porównamy go z tym modelem. Biorąc pod uwagę to, że wiek jest zmienną ilościową oraz wyniki analiz rozkładu wieku w próbce z poprzedniej pracy domowej zdecydowaliśmy się na podział na cztery grupy: <40, 40-60, 60-70, >70.

```
## Call:
## coxph(formula = Surv(dftime, dfstat) ~ hgb + clinstg + chemo +
##       strata(age2), data = df)
##
##          coef exp(coef)  se(coef)      z      p
## hgb      0.004910 1.004923 0.003892 1.262 0.207013
## clinstg  0.443770 1.558572 0.125682 3.531 0.000414
## chemo    -0.549803 0.577063 0.172077 -3.195 0.001398
##
## Likelihood ratio test=20.06 on 3 df, p=0.0001649
## n= 517, number of events= 324
```

Na podstawie wyników modelu nie jesteśmy w stanie powiedzieć, że hemoglobina ma istotny wpływ na czas przeżycia. Pozostałe zmienne mają wpływ. Sprawdźmy, czy założenia są spełnione dla nowego modelu. Ponownie robimy test Schoenfelda z transformacją identycznościową czasu, która wynika z wykresów skalowanych reszt Schoenfelda.

```
##          chisq df      p
## hgb      2.06  1 0.15
## clinstg  1.12  1 0.29
## chemo    1.19  1 0.27
## GLOBAL   4.47  3 0.22
```

Test pokazuje, że możemy stosować model PH. Sprawdźmy dopasowanie modelu z hemoglobiną, stopniem zaawansowania i chemioterapią z warstwami po wieku oraz dopasowanie modelu ze wszystkimi zmiennymi za pomocą wykresu reszt dewiancji.



Na wykresie reszt dewiancji widzimy, że lepiej dopasowany jest model warstwowy. Jednak rozrzut wartości nie przypomina do końca rozkładu normalnego, na przykład nie jest symetryczny wokół 0. Dodatnie wartości reszt dewiancji są bardziej rozrzucone niż ujemne, widać też pojedyncze skupiska wokół innych wartości niż 0. Oznacza to, że model nie jest idealnie dopasowany. Pomimo tego do ostatecznej interpretacji wybieramy model z warstwami.

Z przyjętego przez nas modelu wynika, że zwiększenie stopnia zaawansowania choroby z I na II zwiększa hazard nawrotu choroby 1.558572 razy. W takim razie większy stopień zaawansowania choroby wpływa negatywnie na rokowanie pacjenta. Zastosowanie chemioterapii zmniejsza hazard nawrotu choroby 0.577063 razy. Zatem chemioterapia wpływa pozytywnie na pacjenta. Zwiększenie poziomu hemoglobiny o 1 g/l powoduje zwiększenie hazardu 1.004923 razy. Z testu Walda wynika, że hemoglobina nie ma istotnego wpływu na czas przeżycia, moglibyśmy więc nawet pominąć ją w analizie. Za pomocą tego modelu PH niestety nie jesteśmy w stanie stwierdzić jak wpływa wiek na czas przeżycia bez nawrotu choroby.

```

library(foreign);library(survival);library(rms)
df <- read.csv("C:/Users//PC/Desktop/Studia/BIO//PD1//follic_short_cr.csv")
#-----postać funkcyjne zmiennych ilościowych - reszty martyngalowe-----
chloniak.PH <- coxph(Surv(dftime,dfstat)-1, data=df) #model pusty
mart <- resid(chloniak.PH)
#wiek = funkcja liniowa
plot(df$age , mart, main="Wykres reszt martyngalowych dla age",
      xlab="wiek", ylab="Reszty martyngalowe", cex=0.8, cex.axis=0.5, cex.lab=0.7, cex.main=0.8)
lines(lowess(df$age , mart, iter=0, f=0.6), col="red")

par(mfrow = c(1,2))
# hgb
plot(df$hgb , mart, main="Wykres reszt martyngalowych dla hgb",
      xlab="hgb", ylab="Reszty martyngalowe", cex=0.5, cex.axis=0.5, cex.lab=0.7, cex.main=0.8)
lines(lowess(df$hgb , mart, iter=0, f=0.6), col="red")
#widać obserwacje odstające...
# hgb bez obserwacji odstającej - funkcja liniowa/nawet stała
df2<-df[df$hgb<40, ]
chloniak2.PH <- coxph(Surv(dftime,dfstat)-1, data=df2)
mart2 <- resid(chloniak2.PH)
plot(df2$hgb , mart2, main="Wykres reszt martyngalowych dla hgb",
      xlab="hgb", ylab="Reszty martyngalowe", cex=0.5, cex.axis=0.5, cex.lab=0.7, cex.main=0.8)
lines(lowess(df2$hgb , mart2, iter=0, f=0.6), col="red")

#-----dopasowanie pełnego modelu-----
chloniak.PH.pelny <- coxph(Surv(dftime,dfstat)-age*hgb*clinstg*chemo, data=df)
print(chloniak.PH.pelny)
chloniak2.PH.pelny <- coxph(Surv(dftime,dfstat)-age*hgb*clinstg*chemo, data=df2)
print(chloniak2.PH.pelny)
#-----sprawdzanie założen-----
#krzywa KM - chemicterapia
chloniak.KM.chemo <- survfit(Surv(dftime,dfstat) ~ chemo, data=df)
plot(chloniak.KM.chemo, col=c("red","blue"), xlab="months",ylab="survival probability")
#transformacja
plot(chloniak.KM.chemo, col=c("red","blue"),fun=function(x) log(-log(x)), log="x", firstx=1)

#krzywa KM - st. zawałowoawiania
chloniak.KM.clinstg <- survfit(Surv(dftime,dfstat) ~ clinstg, data=df)
plot(chloniak.KM.clinstg, col=c("red","blue"), xlab="months",ylab="survival probability") #przeciecie /
#transformacja
plot(chloniak.KM.clinstg, col=c("red","blue"),fun=function(x) log(-log(x)), log="x", firstx=1) #przeciecie

#test Schoenfelda
chloniak.PHfit.pelny <- cox.zph(chloniak.PH.pelny, transform="identity")
print(chloniak.PHfit.pelny)
#wykres skalowanych reszt Schoenfelda
par(mfrow=(2,2))
plot(chloniak.PHfit.pelny, df=4, nsmo=10, se=TRUE, var=1)
abline(chloniak.PH.pelny$coeff[1], 0, lty=3)
plot(chloniak.PHfit.pelny, df=4, nsmo=10, se=TRUE, var=2)
abline(chloniak.PH.pelny$coeff[2], 0, lty=3)
plot(chloniak.PHfit.pelny, df=4, nsmo=10, se=TRUE, var=3)
abline(chloniak.PH.pelny$coeff[3], 0, lty=3)
plot(chloniak.PHfit.pelny, df=4, nsmo=10, se=TRUE, var=4)
abline(chloniak.PH.pelny$coeff[4], 0, lty=3)

#WARSTWY ze względem na wiek
df$age2<- cut(df$age,breaks = c(0, 40, 60, 70, 90),labels = c(1, 2, 3, 4), include.lowest = TRUE)
df$age2 <- as.numeric(df$age2)
chloniak.strPH.wiek <- coxph(Surv(dftime,dfstat)-hgb*clinstg*chemo + strata(age2), data=df)
print(chloniak.strPH.wiek)
#-----sprawdzanie założen-----
#test Schoenfelda
chloniak.strPHfit.wiek <- cox.zph(chloniak.strPH.wiek, transform="identity")
print(chloniak.strPHfit.wiek)
#założenia są spełnione
#wykres skalowanych reszt Schoenfelda
plot(chloniak.strPHfit.wiek, df=4, nsmo=10, se=TRUE, var=1)
abline(chloniak.strPH.wiek$coeff[1], 0, lty=3)
plot(chloniak.strPHfit.wiek, df=4, nsmo=10, se=TRUE, var=2)
abline(chloniak.strPH.wiek$coeff[2], 0, lty=3)
plot(chloniak.strPHfit.wiek, df=4, nsmo=10, se=TRUE, var=3)
abline(chloniak.strPH.wiek$coeff[3], 0, lty=3)
#ocena dopasowania modelu za pomocą deviancji
chloniak.devres.strPH.wiek <- residuals(chloniak.strPH.wiek,type="deviance")
chloniak.fitval.strPH.wiek <- predict(chloniak.strPH.wiek,type="lp")
#wykres
par(mfrow=c(1,1))

plot(chloniak.fitval.strPH.wiek,chloniak.devres.strPH.wiek)

# -----ocena modelu pełnego za pomocą deviancji
chloniak.pelny.devres <- residuals(chloniak.PH.pelny,type="deviance")
chloniak.pelny.fitval <- predict(chloniak.PH.pelny,type="lp")
#wykres
plot(chloniak.pelny.fitval,chloniak.pelny.devres)

```