BIOSTATYSTYKA – PROJEKT KOŃCOWY

MARTA SOMMER – BSMAD

23 czerwca 2014

Analizujemy dane dotyczące 927 noworodków, które były karmione piersią przez matki. Interesuje nas, która ze zmiennych charakteryzujących matkę ma wpływ na czas do odstawienia dziecka od piersi.

Wstępna analiza danych

Nasz zbiór danych zawiera następujące zmienne:

`feed` – czas trwania karmienia piersią (w tygodniach)

`koniec_karmienia` – wskaźnik odstawienia od piersi (0 – nie, 1 – tak)

`race` – rasa matki (1 – biała, 2 – czarna, 3 – inna)

`econ` – wskaźnik sytuacji ekonomicznej matki (0 – dobra, 1 – zła)

`smok` – czy matka paliła w czasie ciąży (0 – nie, 1 – tak)

`alco` – czy matka piła alkohol w czasie ciąży (0 – nie, 1 – tak)

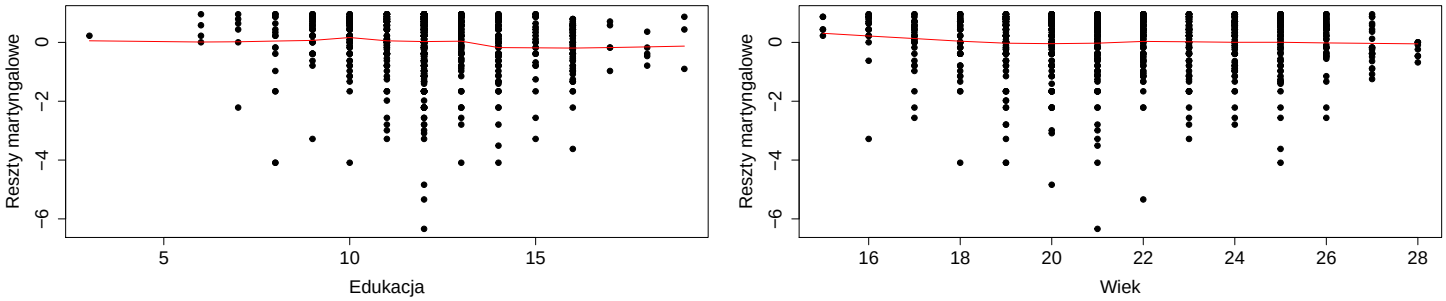
`age` – wiek matki w chwili narodzin dziecka (w latach)

`educ` – wykształcenie matki (lata nauki)

`care` – czy matka korzystała z opieki prenatalnej na początku ciąży (0 – nie, 1 – tak)

Żeby móc przeprowadzić analizę, musimy odpowiednio dostosować nasz zbiór danych. Przede wszystkim zmienną `race` zmienimy na dwie zmienne indykatorowe, gdyż nie powinna ona reprezentować porządku, co robi przy aktualnym kodowaniu. Dostaniemy w ten sposób dwie nowe zmienne `rasa_biała` (równą 1 dla rasy białej i 0 w przeciwnym przypadku) oraz `rasa_czarna` (równą 1 dla rasy czarnej i 0 w przeciwnym przypadku).

Zmienna `age` oraz `educ` są w pewnym przybliżeniu zmiennymi ciągłymi. Sprawdzę więc, czy może istnieje dla nich jakaś odpowiednia postać funkcyjna. W tym celu zbuduję pusty model PH i narysuję jego reszty martyngałowe w zależności od naszych dwóch zmiennych ciągłych. Oto rezultaty:



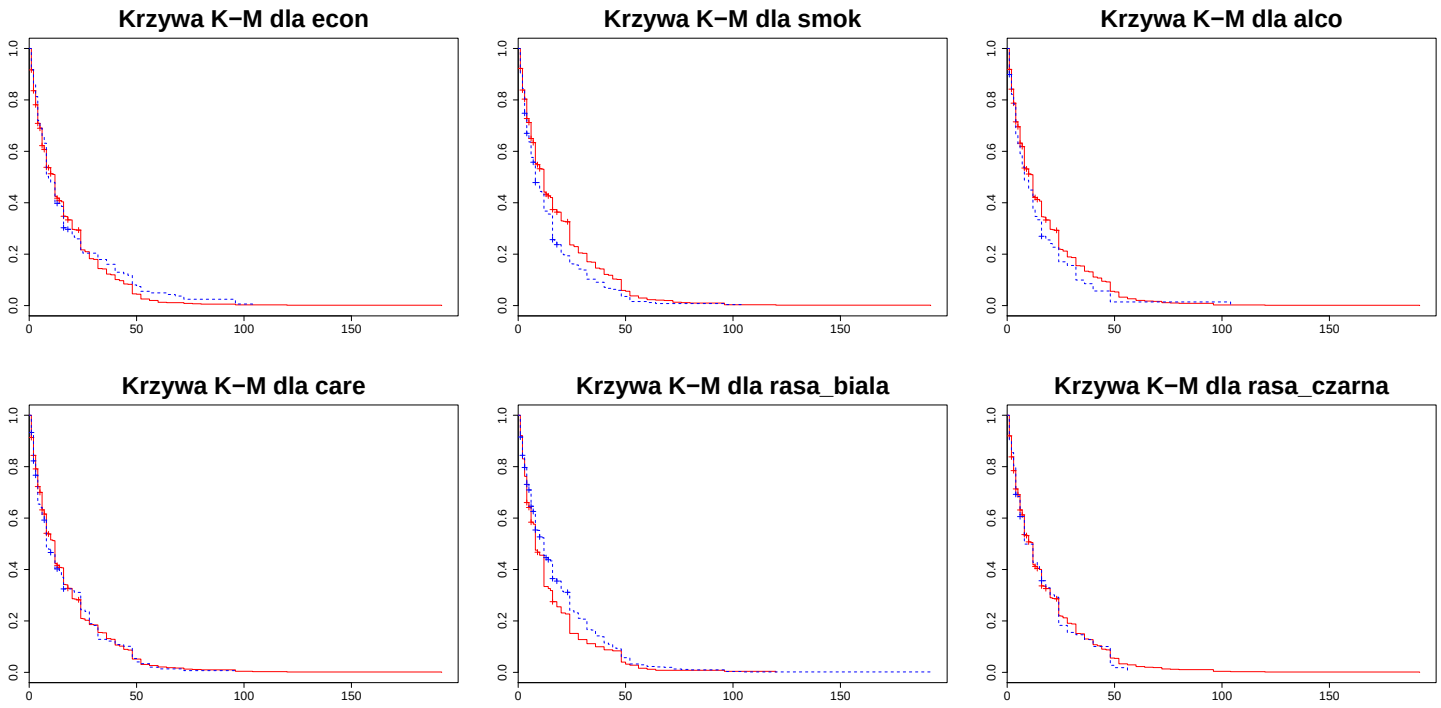
Powyższe wykresy to w przybliżeniu funkcje stałe, tak więc nie ma potrzeby nadawania zmiennym `age` i `educ` żadnej postaci funkcyjnej.

Nasze dane w ostatecznej formie wyglądają więc w następujący sposób:

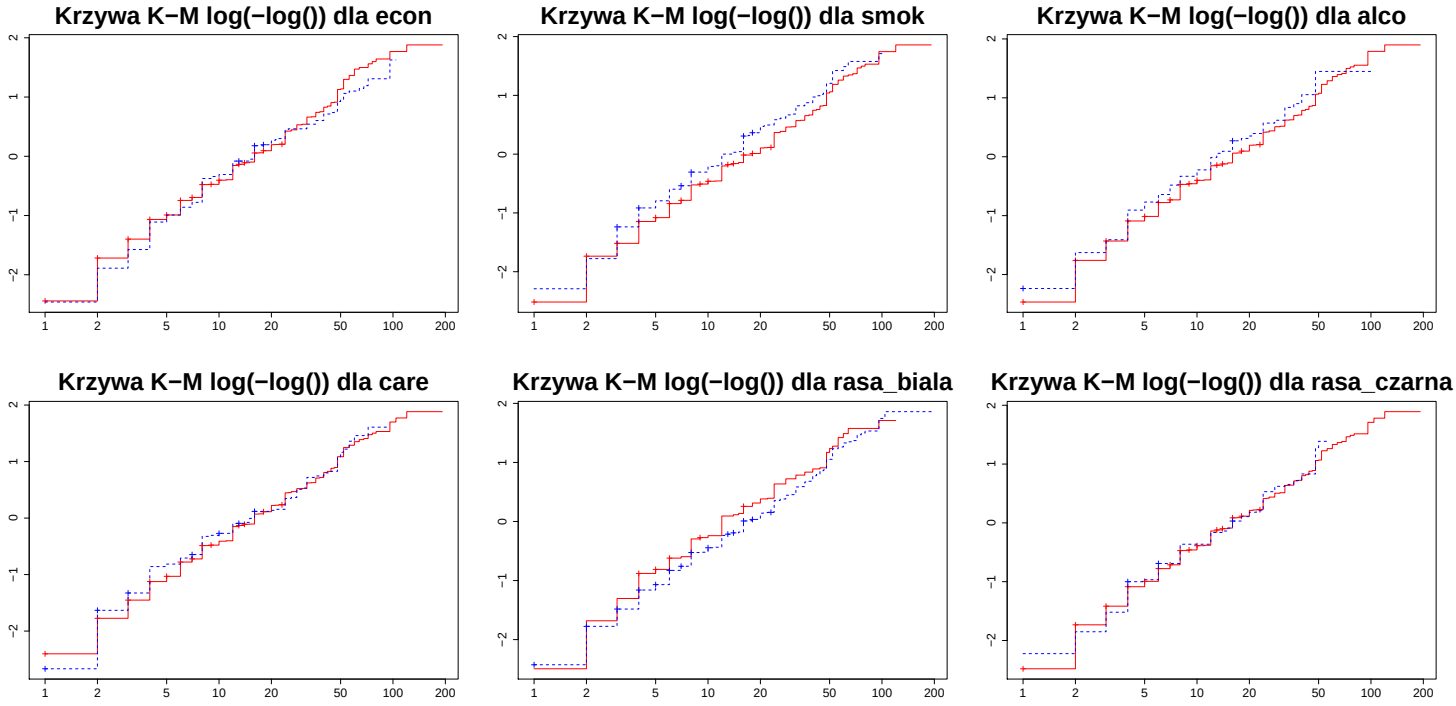
```
  feed koniec_karmienia econ smok alco age educ care rasa_biala rasa_czarna
1   16                1    0    0    1  24   14    0          1           0
2    1                1    0    1    0  26   12    0          1           0
```

Sprawdzenie założeń modelu PH

Na początek narysujmy krzywe Kaplana-Meiera, żeby zobaczyć, które zmienne dobrze różnicują czas odstawienia dziecka od piersi (oczywiście robimy to tylko dla zmiennych dyskretnych).



Widać, że żadna ze zmiennych wyraźnie nie różnicuje czasu do odstawienia dziecka od piersi. Krzywe Kaplana-Meiera co prawda przecinają się (a nie powinny, gdy są spełnione założenia proporcjonalnych hazardów), jednak widać, że te krzywe są po prostu niemal identyczne, dlatego nachodzą na siebie. Przyjrzyjmy się jeszcze ich przekształceniu $log(-log)$.

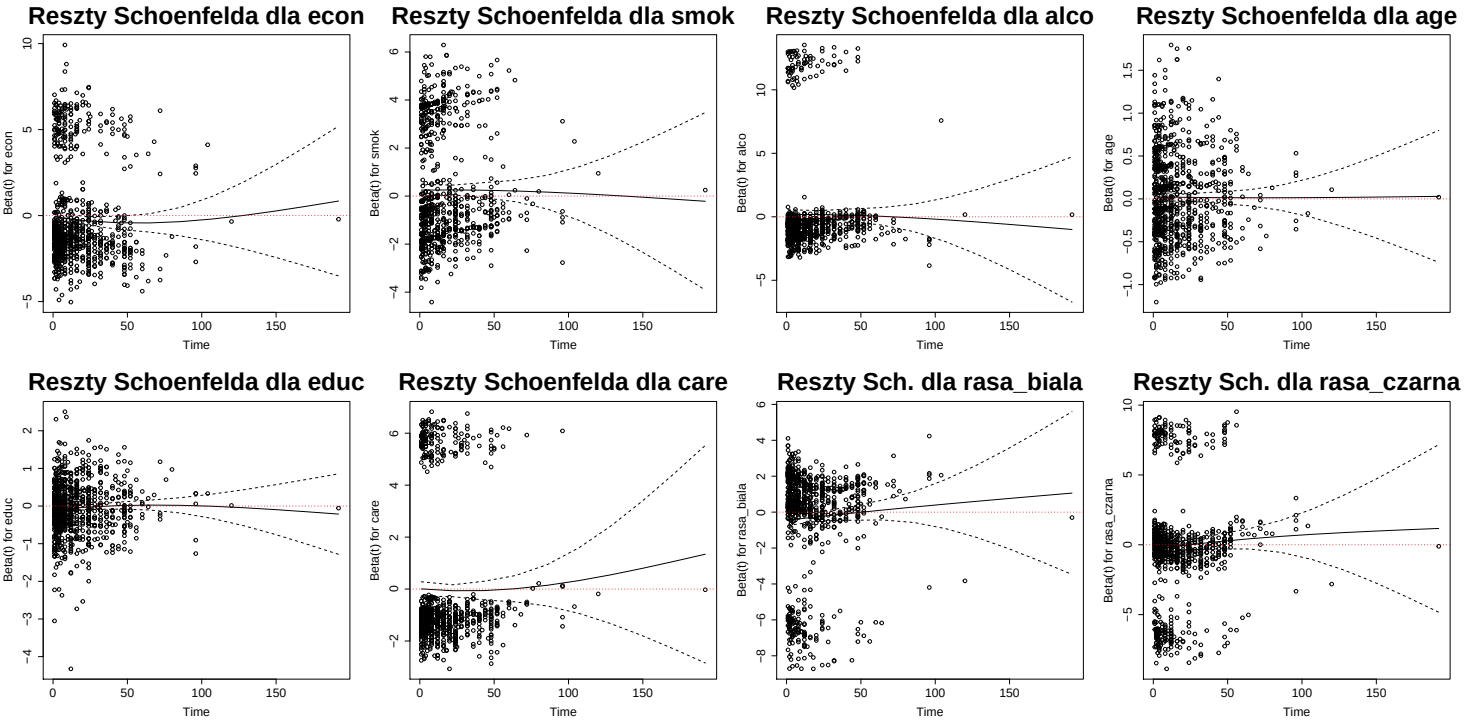


Krzywe powinny być do siebie równoległe. Widać, że nieraz się nawet przecianją. Wydaje się jednak, że zachowują się w miarę równolegle. Jako, że wykresy przysparzają nam nieco trudności w interpretacji, przyjrzyjmy się formalnemu testowi skalowanych reszt Schoenfelda:

```
                rho  chisq     p
econ        -0.02938 0.8025 0.370
smok        -0.00574 0.0296 0.863
alco        -0.01336 0.1616 0.688
age         -0.00519 0.0250 0.874
educ         0.04540 1.8418 0.175
care         0.00526 0.0249 0.875
rasa_biala   0.05437 2.6581 0.103
rasa_czarna  0.05274 2.5210 0.112
GLOBAL            NA 9.8189 0.278
```

P-value każdego z testów jest większe niż $0,05$, zatem nie mamy podstaw do odrzucenia hipotezy, że założnie proporcjonalnych hazardów jest spełnione (również dla zminnych `age` i `educ`, dla których nie mogliśmy narysować krzywych przeżycia Kaplana-Meiera ze względu na ich ciągły charakter). To samo tyczy się tesu globalnego, którego p-value też jest duże i wynosi około $0,28$.

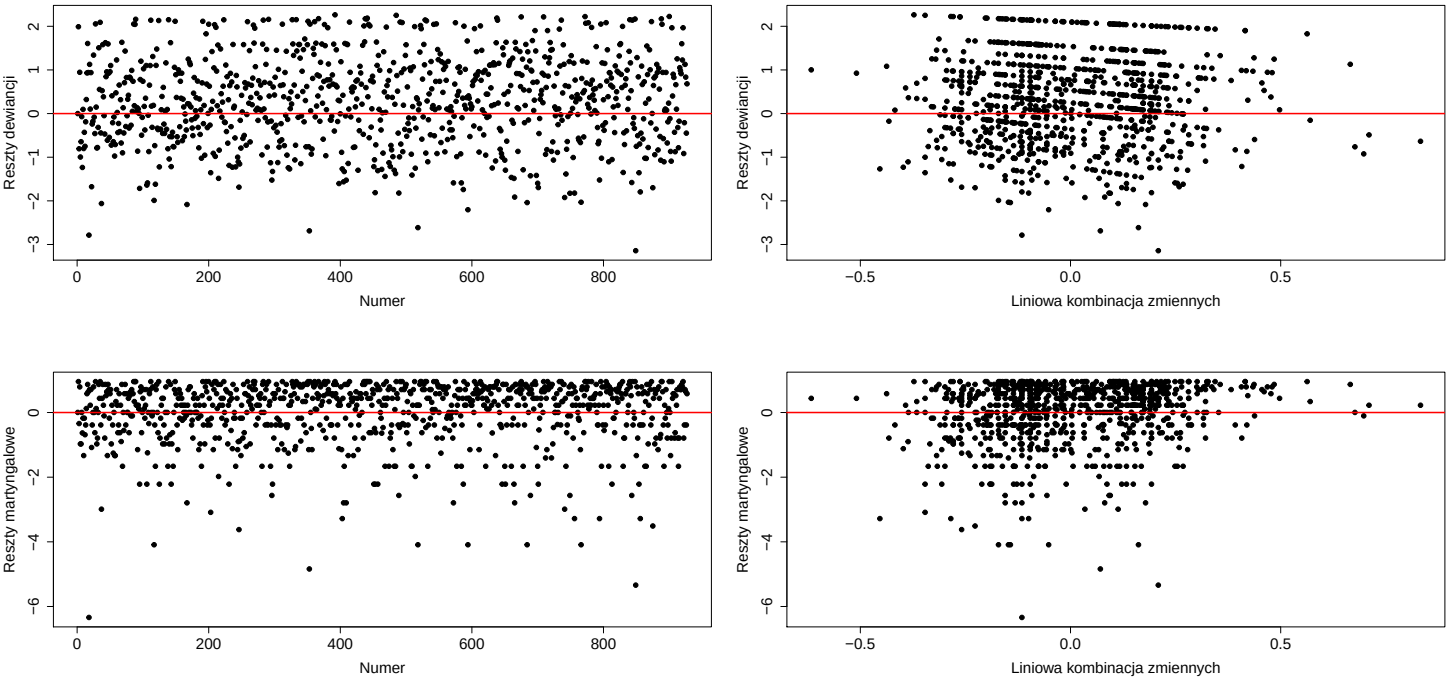
Spójrzmy jeszcze ostatecznie na wykresy reszt Schoenfelda dla kolejnych zmiennych.



Gdy spełnione jest założenie proporcjonalnych hazardów, oczekiwalibyśmy poziomej linii mieszczącej się w pasach ufności. I rzeczywiście, dla każdej zmiennej, wykres jest bliski funkcji stałej. Ostateczny wniostek jest więc taki, że dla żadnej zmiennej objaśniającej nie mamy podstaw uważać, że założenie proporcjonalnych hazardów nie jest spełnione. Przejdźmy więc do dopasowania i analizy modelu PH.

Dopasowanie modelu

Przeanalizujmy teraz ogóle dopasowanie modelu. Przyjrzyjmy się w tym celu czterem wykresom:



Na każdym z wykresów reszty są w miarę symetrycznie rozrzucone wokół prostej $y = 0$. Co więcej, wykresy reszt dewiancji mieszczą się w przedziale $[-3, 3]$, co jest ich dobrą cechą. Niestety reszty martyngałowe nie zachowują się już tak dobrze. Widać wyraźnie, że występuje kilka obserwacji odstających.

Analiza modelu i wnioski

Przyjrzyjmy się współczynnikom z dopasowanego modelu PH:

```
Call:
coxph(formula = Surv(feed, koniec_karmienia) ~ ., data = bb)

  n= 927, number of events= 892

               coef exp(coef) se(coef)     z Pr(>|z|)
econ        -0.2105    0.8102   0.0934 -2.25   0.0243 *
smok         0.2488    1.2824   0.0793  3.14   0.0017 **
alco         0.1682    1.1832   0.1227  1.37   0.1705
age          0.0198    1.0200   0.0165  1.20   0.2297
educ        -0.0557    0.9458   0.0230 -2.43   0.0153 *
care        -0.0265    0.9738   0.0899 -0.30   0.7678
rasa_biala  -0.3047    0.7374   0.0972 -3.13   0.0017 **
rasa_czarna -0.1104    0.8955   0.1287 -0.86   0.3910
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

            exp(coef) exp(-coef) lower .95 upper .95
econ            0.810      1.234     0.675     0.973
smok            1.282      0.780     1.098     1.498
alco            1.183      0.845     0.930     1.505
age             1.020      0.980     0.988     1.054
educ            0.946      1.057     0.904     0.989
care            0.974      1.027     0.817     1.161
rasa_biala      0.737      1.356     0.609     0.892
rasa_czarna     0.895      1.117     0.696     1.152

Concordance= 0.567  (se = 0.012 )
Rsquare= 0.033   (max possible= 1 )
Likelihood ratio test= 31.2  on 8 df,   p=0.00013
Wald test            = 32  on 8 df,   p=9.14e-05
Score (logrank) test = 32.1  on 8 df,   p=9.1e-05
```

Z testu Walda możemy odczytać, które zmienne są istotne w modelu. Mianowicie są to zmienne `econ`, `smok`, `educ` i `rasa_biala`.

Wnioski są zatem następujące. Hazard (w naszym przypadku jest to ryzyko odstawienia dziecka od piersi) jest większy $1,28$ razy, gdy kobieta paliła papierosy w czasie ciąży, $1,18$ razy, gdy piła alkohol, $1,02$ razy, gdy kobieta jest o rok starsza. Pozostałe zmienne hazard zmniejszają. Tak więc, ryzyko odstawienia dziecka od piersi jest $0,81$ razy mniejsze, gdy kobieta ma gorszą sytuację ekonomiczną, $0,94$ razy mniejsze, gdy ma o jeden rok edukacji więcej, $0,97$ razy mniejsze, gdy potrzebowała opieki prenatalnej w pierwszych miesiącach ciąży, $0,74$ razy mniejsze, gdy jest rasy białej i wreszcie $0,9$ razy mniejsze, gdy jest rasy czarnej.

Największy wpływ na zmienę hazardu ma zatem to, czy kobieta jest rasy białej oraz czy w czasie ciąży paliła papierosy.

Z testu największej wiarogodności widać też, że model jako całość jest dobrze dopasowany (p-value równe $0,00013$).

Kod R-owy

```
library("foreign")
library("survival")
b <- read.dta("C:\\Users\\Marta\\Desktop\\Marta\\studia\\rok4\\Biostatystyka\\projekt\\BreastFeeding.dta")
names(b)[2] <- "koniec_karmienia"
n <- nrow(b)
rasa_biala <- ifelse(b$race==1,1,0)
```

```
rasa_czarna <- ifelse(b$race==2,1,0)
bb <- cbind(b[,-3],rasa_biala,rasa_czarna)

modelpusty <- coxph(Surv(feed, koniec_karmienia) ~ 1, data = bb)
mart <- resid(modelpusty)
plot(bb$educ, mart, xlab="Edukacja", ylab="Reszty martyngałowe",pch=19, cex.main=3,cex.axis=1.5,cex.lab=1.5)
lines(lowess(bb$educ, mart, iter=0, f=0.6), col="red")
plot(bb$age, mart, xlab="Wiek", ylab="Reszty martyngałowe", pch=19, cex.main=3,cex.axis=1.5,cex.lab=1.5)
lines(lowess(bb$age, mart, iter=0, f=0.6), col="red")

km1 <- survfit(Surv(feed,koniec_karmienia) ~ econ, data=bb, conf.type="none")
plot(km1,col=c("red","blue"), lty=1:2, main="Krzywa K-M dla econ", cex.main=3,cex.axis=1.5)
km2 <- survfit(Surv(feed,koniec_karmienia) ~ smok, data=bb, conf.type="none")
plot(km2,col=c("red","blue"), lty=1:2, main="Krzywa K-M dla smok", cex.main=3,cex.axis=1.5)
km3 <- survfit(Surv(feed,koniec_karmienia) ~ alco, data=bb, conf.type="none")
plot(km3,col=c("red","blue"), lty=1:2, main="Krzywa K-M dla alco", cex.main=3,cex.axis=1.5)
km4 <- survfit(Surv(feed,koniec_karmienia) ~ care, data=bb, conf.type="none")
plot(km4,col=c("red","blue"), lty=1:2, main="Krzywa K-M dla care", cex.main=3,cex.axis=1.5)
km5 <- survfit(Surv(feed,koniec_karmienia) ~ rasa_biala, data=bb, conf.type="none")
plot(km5,col=c("red","blue"), lty=1:2, main="Krzywa K-M dla rasa_biala", cex.main=3,cex.axis=1.5)
km6 <- survfit(Surv(feed,koniec_karmienia) ~ rasa_czarna, data=bb, conf.type="none")
plot(km6,col=c("red","blue"), lty=1:2, , main="Krzywa K-M dla rasa_czarna", cex.main=3,cex.axis=1.5)

plot(km1,col=c("red","blue"), lty=1:2, fun=function(x) log(-log(x)), log="x", firstx=1, main="Krzywa K-M log(-log()) dla e
plot(km2,col=c("red","blue"), lty=1:2, fun=function(x) log(-log(x)), log="x", firstx=1, main="Krzywa K-M log(-log()) dla s
plot(km3,col=c("red","blue"), lty=1:2, fun=function(x) log(-log(x)), log="x", firstx=1, main="Krzywa K-M log(-log()) dla a
plot(km4,col=c("red","blue"), lty=1:2, fun=function(x) log(-log(x)), log="x", firstx=1, main="Krzywa K-M log(-log()) dla c
plot(km5,col=c("red","blue"), lty=1:2, fun=function(x) log(-log(x)), log="x", firstx=1, main="Krzywa K-M log(-log()) dla r
plot(km6,col=c("red","blue"), lty=1:2, fun=function(x) log(-log(x)), log="x", firstx=1, main="Krzywa K-M log(-log()) dla r

ph <- coxph(Surv(feed, koniec_karmienia)~., data=bb)
test <- cox.zph(ph, transform="identity")

plot(test, df=3, nsmo=10, se=TRUE, var=1, main="Reszty Schoenfelda dla econ", pch=19,cex.main=3,cex.axis=1.5,cex.lab=1.5)
abline(0, 0, lty=3, col="red")
plot(test, df=3, nsmo=10, se=TRUE, var=2, main="Reszty Schoenfelda dla smok", pch=19,cex.main=3,cex.axis=1.5,cex.lab=1.5)
abline(0, 0, lty=3, col="red")
plot(test, df=3, nsmo=10, se=TRUE, var=3, main="Reszty Schoenfelda dla alco", pch=19,cex.main=3,cex.axis=1.5,cex.lab=1.5)
abline(0, 0, lty=3, col="red")
plot(test, df=3, nsmo=10, se=TRUE, var=4, main="Reszty Schoenfelda dla age", pch=19,cex.main=3,cex.axis=1.5,cex.lab=1.5)
abline(0, 0, lty=3, col="red")
plot(test, df=3, nsmo=10, se=TRUE, var=5, main="Reszty Schoenfelda dla educ", pch=19,cex.main=3,cex.axis=1.5,cex.lab=1.5)
abline(0, 0, lty=3, col="red")
plot(test, df=3, nsmo=10, se=TRUE, var=6, main="Reszty Schoenfelda dla care", pch=19,cex.main=3,cex.axis=1.5,cex.lab=1.5)
abline(0, 0, lty=3, col="red")
plot(test, df=3, nsmo=10, se=TRUE, var=7, main="Reszty Sch. dla rasa_biala", pch=19,cex.main=3,cex.axis=1.5,cex.lab=1.5)
abline(0, 0, lty=3, col="red")
plot(test, df=3, nsmo=10, se=TRUE, var=8, main="Reszty Sch. dla rasa_czarna", pch=19,cex.main=3,cex.axis=1.5,cex.lab=1.5)
abline(0, 0, lty=3, col="red")

dewiancja <- residuals(ph, type="deviance")
coef <- ph$linear.predictors
plot(1:n, dewiancja, pch=19, xlab="Numer", ylab="Reszty dewiancji",cex.main=3,cex.axis=1.5,cex.lab=1.5)
abline(0,0, col="red", lwd=2)
plot(coef, dewiancja, pch=19, xlab="Liniowa kombinacja zmiennych", ylab="Reszty dewiancji",cex.main=3,cex.axis=1.5,cex.lab
abline(0,0, col="red", lwd=2)
plot(1:n, mart, pch=19, xlab="Numer", ylab="Reszty martyngałowe",cex.main=3,cex.axis=1.5,cex.lab=1.5)
abline(0,0, col="red", lwd=2)
plot(coef, mart, pch=19, xlab="Liniowa kombinacja zmiennych", ylab="Reszty martyngałowe",cex.main=3,cex.axis=1.5,cex.lab=1
abline(0,0, col="red", lwd=2)
```