Sesja komputerowa 5: R

1. Dane dotyczące AIDS i SI z artykułu Puttera i kolegów znajdują się w pliku **aidssi.csv**. By je wczytać, używamy funkcji **read.csv**:

```
dane <- read.csv(file="t:\\burzykowskit\\biostat\\datasets\\aidssi.csv")
```

Obiekt (dataframe) **dane** zawiera pięć zmiennych: *patnr* (numer chorego); *time* (czas obserwacji w latach); *status* jest wskaźnikiem zdarzenia (0 = cenzurowanie, 1 = AIDS, 2 = SI); *cause* jest zmienną tekstową z opisem zdarzenia (event-free, AIDS, lub SI); *ccr5* jest wskaźnikiem delecji w genie CCR5 Δ32 (WW = brak delecji (wilde-type), WM = delecja).

Dane zawierają pięć obserwacji bez informacji o delecji w genie CCR5 Δ32. Nie będziemy ich w analizach używać, więc usuniemy je z danych:

```
dane.red <- subset(dane, ccr5 %in% c("WW","WM"))
```

Dodatkowo, potrzebować będziemy zmiennej **event**, która wskazuje zdarzenia niezależnie od ich typu:

```
dane.red <- data.frame(dane.red,event=ifelse(dane.red$status>0,1,0))
```

Będziemy również potrzebować będziemy zmiennej **si**, kodującej SI jako zdarzenie nas interesujące, (0 = cenzurowanie, 1 = AIDS, 2 = SI):

```
dane.red <- data.frame(dane.red, si=3-dane.red$status-3*(
dane.red$status==0))
```

Od tej pory będziemy używać zmodyfikowanych danych:

```
attach(dane.red)
```

2. Dla celów analizy będziemy używać funkcji z pakietu *cmprisk*. Inicjalizujemy go wykonując polecenie:

```
library(cmprsk)
```

Wywołanie polecenia autmoatycznie uruchamia rónież pakiet *survival*.

3. Najpierw użyjemy metody Kaplana-Meiera dla AIDS i SI. W tym celu używamy znanej już nam finkcji **survfit()**:

```
aids.km <- survfit(Surv(time, status==1) ~ 1, data = dane.red)
si.km <- survfit(Surv(time, status==2) ~ 1, data = dane.red)
```

Zwróćmy uwagę, że w definicji obiektu **Surv** w pierwszym poleceniu używamy warunku **status==1**. Uzyskujemy w ten sposób wskaźnik zdarzeń równy 1 dla AIDS i 0 dla cenzurowania i SI. W drugim poleceniu w analogiczny sposób uzyskujemy wskaźnik zdarzeń równy 1 dla SI i 0 dla cenzurowania i AIDS.

Następnie konstruujemy wykres oszacowanej funkcji dla AIDS i dopełnienia funkcji dla SI:

```
plot(aids.km, mark.time=FALSE, conf.int=FALSE, col=2, xscale=1, xmax=13,
xlab="Years post-HIV infection")
lines(si.km, fun="event", mark.time=FALSE, conf.int=FALSE, col=3, xmax=13)
```

```
text(8, .8, "KM, AIDS", col=2)
text(8, .2,"1-KM, SI", col=3)
```

Polecenie **plot()** wykreśla oszacowaną funkcję dla AIDS. Polecenie **lines()** dodaje wykres dopełnienia funkcji dla SI. Istotną rolę w tym celu gra argument **fun="event"**, który wskazuje, że chcemy wykreślić dopełnienie oszacowanej funkcji. Uzyskany wykres odpowiada rycinie ze slajdu 20.

3. Oszacujemy teraz sub-dystrybuanty (funkcje skumulowanej częstości) dla AIDS i SI. W tym celu również możemy posłużyć się finkcją **survfit()**, ale w zmodyfikowanej postaci:

```
ci <- survfit(Surv(time, status, type=”mstate”) ~ 1, data = dane.red)
```

W szczególności, w definicji obiektu **Surv** używamy wskaźnika **status**, który przyjmuje wartość 0 dla obserwacji cenzurowanych oraz 1 dla AIDS i 2 dla SI. Użycie argumentu **type=”mstate”** powoduje, że wskaźnik jest traktowany jako czynnik dla którego pierwszy poziom (0) jest identyfikatorem obserwacji cenzurowanych, pozostałe poziomy identyfikują konkurencyjne ryzyka.

Następnie konstruujemy wykres oszacowanych sub-dystrybuant dla AIDS i SI:

```
plot(ci, mark.time=FALSE, conf.int=FALSE, col=(2,3), xscale=1, xmax=13,
xlab="Years post-HIV infection")
text(8, .4, "AIDS", col=2)
text(8, .2, "SI", col=3)
```

4. Sub-dystrybuanty możemy również oszacować używając funkcji **cuminc()** z pakietu *cmprsk*:

```
ci.cmprsk <- cuminc(ftime=time,fstatus=status)
```

Argument **ftime** definiuje zmienną zawierającą czas obserwacji, **fstatus** definiuje wskaźnik zdarzeń (domyślnie 0 oznacza obserwacje cenzurowane). W istocie, możnaby użyć skróconej formy tego polecenia, a mianowicie **cuminc(time, status)**.

Podstawowe wyniki i oszacowania zawarte w obiekcie **ci.cmprsk** możemy uzyskać przy użyciu metody **print.cuminc()**:

```
print(ci.cmprsk,ntp=2)
```

Argument **ntp** wskazuje liczbę chwil czasu, dla których chcemy uzyskać oszacowanie sub-dystrybuant i ich wariancji. Wykres oszacowanych sub-dystrybuant uzyskujemy przy pomocy metody **plot.cuminc()**:

```
plot(ci.cmprsk,lty=2,c=2:3,xlab="Years post-HIV infection", ylab="CIF",
curvlab= c("AIDS","SI"))
```

5. Funkcja **cuminc()** pozwala na porównanie oszacowań sub-dystrybuant przy pomocy testu zaproponowanego przez Gray’a. Zilustrujemy tę możliwość rozważając grupy zdefiniowane delecją genu CCR5:

```
ci.ccr5 <- cuminc(time,status,group=ccr5)
```

Argument **group** wskazuje zmienną definiującą grupy obserwacji, dla których chcemy uzyskać oszacowania subdystrybuant i ich porównanie. Wyniki uzyskujemy przy użyciu metody **print.cuminc()**:

```
print(ci.ccr5,ntp=2)
```

Wyniki testu wskazują na istotną statystycznie różnicę między sub-dystrybuantami AIDS dla WW i WM (p=0.0003) i brak różnicy dla SI (p=0.94).

Wykres oszacowanych sub-dystrybuant uzyskujemy przy pomocy metody **`plot.cuminc()`**:

```
plot(ci.ccr5,curvlab=c("WM,AIDS","WW,AIDS","WM,SI","WW,SI"),
lty=c(1,2,1,2), c=c(2,2,3,3), xlab="Years post-HIV infection", ylab="CIF")
```

6. Uzyskanie wykresu oszacowań sub-dystrybuant dla różnych grup i określonego typu zdarzeń przy użyciu metody **`plot.cuminc()`** wymaga trochę pracy. W tym celu z obiektu klasy *cuminc* należy wybrać odpowiednie składniki.

W szczególności, obiekt klasy *cuminc* jest listą zawierającą tyle składników, ile jest wszsytkich kombinacji typów zdarzeń i grup, plus jeden (jeśli grup jest więcej niż jedna). Ten ostatni składnik jest listą zawierającą wyniki testów porównujących sub-dystrybuanty. Pozostałe składniki są również listami, zawierającymi oszacowania sub-dystrybuant i ich wariancje dla wszystkich grup i zdarzeń.

Informację o strukturze obiektu **`ci.ccr5`** uzyskujemy przy pomocy polecenia

```
str(ci.ccr5)
```

Obiekt jest listą składającą się z pięciu składników, o nazwach: **`WM 1`**, **`WW 1`**, **`WM 2`**, **`WW 2`**, oraz **`Tests`**. Pierwsze cztery składniki są listami ze składnikami: **`time`**, **`est`**, oraz **`var`**. Aby uzyskać wykres oszacowań sub-dystrybuant dla WW i WM dla AIDS, musimy wybrać pierwsze dwie listy składniki. W tym celu możemy użyć następującego polecenia:

```
aids.ccr5 <- list(ci.ccr5$"WM 1", ci.ccr5$"WW 1")
```

Tak utworzoną listę używamy w poleceniu **`plot.cuminc()`**:

```
plot.cuminc(aids.ccr5,curvlab=c("WM","WW"),lty=1:2,c=2,xlab="Years post-HIV
infection",ylab="CIF")
```

Zwróćmy uwagę, że musimy *explicite* zaznaczyć użycie metody **`plot.cuminc()`**. Obiekt **`aids.ccr5`** nie jest bowiem obiektem klasy *cuminc* i użycie polecenia **`plot()`** nie prowadzi do wyboru metody **`plot.cuminc()`**, co skutkuje błędem egzeukcji polecenia.

Wykresy oszacowań sub-dystrybuant dla WW i WM dla SI otrzymujemy w następujący sposób:

```
si.ccr5 <- list(ci.ccr5$"WM 2", ci.ccr5$"WW 2")
plot.cuminc(si.ccr5,curvlab=c("WM","WW"),lty=1:2,c=3,xlab="Years post-HIV
infection",ylab="CIF")
```

Uzyskane wykresy odpowiadają rycinom ze slajdu 30.

7. Uzyskanie wykresu oszacowań sub-dystrybuant dla różnych grup i określonego typu zdarzeń przy użyciu metody **`plot.survfit()`** jest łatwiejsze.

Najpierw szacujemy sub-dystrybuanty:

```
ci.sfit <- survfit(Surv(time, event) ~ ccr5, etype=cause , data = dane.red)
```

A następnie wykreślamy je dla AIDS:

```
plot(ci.sfit, lty=c(1,0,2,0), col=2, mark.time=FALSE, conf.int=FALSE,
xscale=1, xmax=13, xlab="Years post-HIV infection")
```

```
text(6, .275, "WW, AIDS", col=2)
text(6, .075, "WM, AIDS", col=2)
```

Argument **lty=c(1,0,2,0)** pozwala na „wygaszenie” wykresu sub-dystrybuant dla SI.

Podobnie dla SI:

```
plot(ci.sfit, lty=c(0,1,0,2), col=3, mark.time=FALSE, conf.int=FALSE,
xscale=1, xmax=13, xlab="Years post-HIV infection")
text(6, .275, "WW, AIDS", col=3)
text(8, .225, "WM, AIDS", col=3)
```

8. Porównania oszacowań sub-dystrybuant możemy dokonać przy pomocy testu zaproponowanego przez Pepe i Mori (1993). W tym celu musimy użyć funkcji **compCIF()** spoza pakietu *cmprsk*. Dostęp do funkcji uzyskujemy przy pomocy następującego polecenia:

```
source("D:\\Tomek\\Courses\\Przemek\\Materials\\CompetingRisks\\Rprograms\\
compCIF.txt")
```

Test dla sub-dystrybuant dla AIDS uzyskujemy przy użyciu następującego polecenia:

```
compCIF(time,status,ccr5)
```

Pierwszy argument wskazuje zmienną zawierającą czas obserwacji (**time**), drugi - zmienną zawierającą wskaźniki zdarzeń i cenzurowania (**status**). Konwencja wymaga użycia wartości 0 dla obserwacji cenzurowanych, 1 dla interesującego nas zdarzenia, i 2 dla konkurencyjnego ryzyka (ryzyk). Trzeci argument wskazuje zmienną grupującą (**ccr5**).

Wyniki testu wskazują na istotną statystycznie różnicę między sub-dystrybuantami AIDS dla WW i WM (p=0.00005). Wartość statystyki testowej jest większa niż dla testu Gray’a.

Test dla sub-dystrybuant dla SI uzyskujemy przy użyciu następującego polecenia:

```
compCIF(time,si,ccr5)
```

Wyniki testu wskazują na nieistotną statystycznie różnicę między sub-dystrybuantami SI dla WW i WM (p=0.54). Wartość statystyki testowej jest większa niż dla testu Gray’a.

10. Model PH dla funkcji hazardu „specyficznych dla typu” uzyskujemy przy użyciu funkcji **coxph()**:

```
summary(coxph.aids <- coxph(Surv(time, status == 1) ~ ccr5, data =
dane.red))
summary(coxph.si <- coxph(Surv(time, status == 2) ~ ccr5, data = dane.red))
```

Zwróćmy uwagę, że w definicji obiektu **Surv** w pierwszym poleceniu używamy warunku **status==1**. Uzyskujemy w ten sposób wskaźnik zdarzeń równy 1 dla AIDS i 0 dla cenzurowania i SI. W drugim poleceniu w analogiczny sposób definiujemy wskaźnik zdarzeń równy 1 dla SI i 0 dla cenzurowania i AIDS.

Wyniki odpowiadają rezultatom podanym na slajdzie 31 (modulo znak współczynników).

11. Model PH dla funkcji hazardu sub-dystrybuanty uzyskujemy przy użyciu funkcji **crr()** z pakietu *cmprsk*. W tym celu musimy najpierw utworzyć macierz zawierającą wartości zmiennych objaśniających:

```
crr.mat <- model.matrix(~factor(ccr5))[,-1]
```

Zwróćmy uwagę na użycie indeksu **`[,-1]`**. Usuwa on z utworzonej macierzy pierwszą kolumnę, zawierającą wyraz wolny (nie jest on potrzebny w modelu PH).

Następnie używamy poleceń

```
summary(mod.aids <- crr(ftime=time, fstatus=status, cov1=crr.mat, failcode=1))
summary(mod.si <- crr(time,status,crr.mat,failcode=2))
```

Pierwsze dopasowuje model dla AIDS, drugie dla SI. W obu poleceniach pierwszy argument (**`ftime`**) wskazuje na zmienną zawierającą czas obserwacji, a drugi (**`fstatus`**) – zmienną zawierającą wskaźnik zdarzenia. Trzeci argument (**`cov1`**) podaje macierz zawierającą wartości zmiennyuch objaśniających. Czwarty (**`failcode`**) wskazuje kod typ zdarzenia dla zmiennej wskazanej w argumencie **`fstatus`**, dla którego szacowany jest model.

Wyniki odpowiadają rezultatom podanym na slajdzie 34 (modulo znak współczynników).

12. Konstruujemy oszacowania funkcji skumulowanych częstości odpowiadających modelowi PH dla funkcji hazardu sub-dystrybuanty dla AIDS:

```
aids.pred <- predict(mod.aids,cov1=rbind(0,1))
```

Argument **`cov1`** podaje macierz wartości zmiennych objaśniających, dla których sub-dystrybuanty mają być oszacowane. Porządek kolumn tej macierzy powinien odpowiadać porządkowi kolumn w macierzy użytej w argumencie **`cov1`** funkcji **`crr()`**.

Następnie wykreślamy uzyskane oszacowania sub-dystrybuant:

```
plot(aids.pred,lty=1:2,color=2)
text(6, .275, "WW, AIDS", col=2)
text(8, .15, "WM, AIDS", col=2)
```

Podobnego syntaksu używamy dla SI:

```
si.pred <- predict(mod.si,rbind(0,1))
plot(si.pred,lty=1:2,color=3)
text(6, .275, "WM, SI", col=3)
text(8, .225, "WW, SI", col=3)
```

Wykresy odpowiadają rycinom podanym na slajdzie 34.

Sesja komputerowa 5: SAS

1. Dane dotyczące AIDS i SI z artykułu Puttera i kolegów znajdują się w pliku `aidssi.sas7bdat`. Aby uzyskać do nich dostęp, musimy najpierw wskazać katalog, w którym znajduje się plik. W tym celu używamy komendy `libname`:

```
libname pw "t:\burzykowskit\biostat\datasets\";
```

Z danych musimy usunąć chorych bez infromacji o delecji w genie CCR5. Ponadto potrzebujemy dodatkowych zmiennych. W celu ich konstrukcji, używamy następującego syntaksu:

```
data aidssi;
  set pw.aidssi;
  if (ccr5 in ("WM","WW"));
  ccr5_num=(ccr5="WM");
  aids=status;
  si=3-status-3*(status=0);
run;
```

Polecenie `if (ccr5 in ("WM","WW"));` włącza do zbioru `aidssi` tylko obserwacje z wartościami zmiennej `ccr5` równymi `"WM"` lub `"WW"`. Zmienna `ccr5_num` jest numerycznym odpowiednikiem zmiennej tekstowej `ccr5` o wartościach 0 i 1 dla, odpowiednio, `"WW"` i `"WM"`. Zmienna `aids` jest wskaźnikiem zdarzenia równym 0 dla obserwacji cenzurowanych, 1 dla AIDS (traktowanym jako zdarzenie nas interesujące), i 2 dla SI (traktowanego jako konkurencyjne ryzyko). Zmienna `aids` jest wskaźnikiem zdarzenia równym 0 dla obserwacji cenzurowanych, 1 dla SI (traktowanym jako zdarzenie nas interesujące), i 2 dla AIDS (traktowanego jako konkurencyjne ryzyko).

2. Oszacowanie funkcji skumulowanej częstości (sub-dystrybuanty) uzyskujemy przy pomocy makra znajdującego się w pliku `cuminc.sas`. Aby uzyskać do niego dostęp, używamy następującego polecenia:

```
%inc "t:\burzykowskit\biostat\programs\cuminc.sas";
```

Następnie wywołujemy makro dla AIDS:

```
%cuminc(ds=aidssi , time=time , cenvble=status ,interest=1 ,group=ccr5 )     ;
```

Parametr `ds` wskazuje zbiór danych, `time` – zmienną zawierającą czas obserwacji, a `cenvble` – zmienną zawierającą wskaźniki zdarzeń/cenzurowania. Kod zdarzenia nas interesującego identyfikujemy poprzez parametr `interest`. Ostatni parametr, `group`, wskazuje zmienną identyfikującą porównywane grupy. Może to być zmienna numeryczna lub tekstowa.

Wykonanie makro skutkuje uzyskaniem oszacowań sub-dystrybuant dla porównywanych grup. Uzyskujemy również wykres sub-dystrybuant. Dodatkowo konstruowany jest wykres warunkowego prawdopodobieństwa zajścia do zdarzenia nas interesującego przed upływem chwili czasu *t* pod warunkiem niezajścia zdarzenia(ń) konkurencyjnego(ych).

Oszacowania dla SI uzyskujemy w następujący sposób:

```
%cuminc(ds=aidssi , time=time , cenvble=status ,interest=2 ,group=ccr5 )     ;
```

3. Porównanie funkcji skumulowanych częstości (sub-dystrybuant) uzyskujemy przy pomocy makra znajdującego się w pliku `compcif.sas`. Makro to używa testu zaproponowanego przez Pepe i Mori (1993), a nie testu Gray'a.

Aby uzyskać dostęp do makra, używamy następującego polecenia:

```
%inc "t:\burzykowskit\biostat\programs\compcif.sas";
```

Następnie wywołujemy makro dla AIDS:

```
%compcif(ds=aidssi, time=time , cens=aids, group=ccr5_num, val1=0, val2=1);
```

Parametr `ds` wskazuje zbiór danych, `time` – zmienną zawierającą czas obserwacji, a `cenvble` – zmienną zawierającą wskaźniki zdarzeń/cenzurowania. Ta ostatnia musi przyjmować wartość 0 dla obserwacji cenzurowanych, 1 dla interesującego nas zdarzenia, a 2 – dla pozostałych zdarzeń (konkurujących ryzyk). Stąd wskazanie zmiennej `aids`. Parametr `group` wskazuje zmienną (numeryczną) identyfikującą porównywane grupy. Test przeprowadza porównanie dla tylko dwóch grup, których identyfikatory podawane są przy pomocy parametrów `val1` i `val2`.

Wykonanie makro skutkuje uzyskaniem oszacowań sub-dystrybuant dla porównywanych grup. Uzyskujemy również rezultat testu. Wynik testu jest istotny statystycznie.

Porównanie sub-dystrybuant dla SI uzyskujemy w następujący sposób:

```
%compcif(ds=aidssi, time=time , cens=si, group=ccr5_num, val1=0, val2=1);
```

Wynik testu nie jest istotny statystycznie.

4. Model PH dla sub-hazardów dla AIDS dopasowujemy przy użyciu procedury PHREG:

```
proc phreg data=aidssi;
 class ccr5;
 model time*status(0,2)=ccr5 / rl;
run;
```

Polecenie `time*status(0,2)` wskazuje, że jako obserwacje cenzurowane traktujemy obserwacje z wartościami zmiennej `status` równymi 0 (cenzurowanie) i 2 (SI).

Model PH dla sub-hazardów dla SI dopasowujemy odpowiednio modyfikując polecenie procedury PHREG:

```
proc phreg data=aidssi;
 class ccr5;
 model time*status(0,1)=ccr5 / rl;
run;
```

Tym razem polecenie `time*status(0,1)` wskazuje, że jako obserwacje cenzurowane traktujemy obserwacje z wartościami zmiennej `status` równymi 0 (cenzurowanie) i 1 (AIDS).

5. Model PH dla funkcji hazardu dla sub-dystrybuanty dla AIDS dopasowujemy przy użyciu makra znajdującego się w pliku `pshreg.sas`. Aby uzyskać dostęp do makra, używamy następującego polecenia:

```
%inc "t:\burzykowskit\biostat\programs\pshreg.sas";
```

Następnie wywołujemy makro dla AIDS:

```
%pshreg(data=aidssi, time=time, cens=status, failcode=1, cencode=0, varlist=ccr5, class = ccr5, options=rl ties=EFRON);
```

Parametr `data` wskazuje zbiór danych, `time` – zmienną zawierającą czas obserwacji, a `cens` – zmienną zawierającą wskaźniki zdarzeń/cenzurowania. Interesujące nas zdarzenia identyfikowane są kodem podanym przy pomocy parametru `failcode` – w tym przypadku to 1, tzn., AIDS. Parametr `cencode` wskazuje kod identyfikujący obserwacje cenzurowane – w tym przypadku to 0. Pozostałe kody zmiennej wskazanej parametrem `cens` traktowane są jako identyfikatory konkurencyjnych ryzyk. Parametr `varlist` podaje listę zmiennych objaśniających, które mają być użyte w modelu; przy użyciu `class` wskazujemy zmienne czynnikowe. Parametr `options` pozwala na użycie opcji dostępnych dla polecenia `model` w procedurze PHREG; w naszym przypadku to `rl ties=EFRON`.

Model PH dla funkcji hazardu dla sub-dystrybuanty dla SI uzyskujemy przy pomocy następującego wywołania makra:

```
%pshreg(data=aidssi, time=time, cens=status, failcode=2, cencode=0, varlist=ccr5, class = ccr5, options=rl ties=EFRON);
```