Procedury selekcji zmiennych w modelach AFT i Coxa

Analiza przeżycia

Zuzanna Klaman

2026-02-17

Spis treści

1 Dalsza diagnostyka modeli

```
library(survival)
data(lung)
dane <- lung[, c("time", "status", "age", "sex", "ph.ecog", "ph.karno")]
dane <- na.omit(dane)
dane$status <- ifelse(dane$status == 2, 1, 0)
srednia_wieku <- mean(dane$age)
srednia_karno <- mean(dane$ph.karno)
dane$age <- dane$age - srednia_wieku
dane$ph.karno <- dane$ph.karno - srednia_karno
```

1.1 Procedury wyboru zmiennych i optymalizacja modelu AFT z użyciem kryteriów AIC i BIC

Rozważono model przyspieszonego czasu awarii, w którym bazowa funkcja przeżycia opisana jest rozkładem Weibulla. Jako zmienną zależną przyjęto zmienną `time`, natomiast do opisu charakterystyk pacjentów wykorzystano zmienne `age`, `sex`, `ph.ecog` oraz `ph.karno`.

Podpunkt (a)

W celu doboru zmiennych do modelu `AFT` zastosujemy metodę eliminacji, opartą na wynikach testu ilorazu wiarygodności. Na początku zbudowano pełny model i w każdym kroku eliminowano zmienną z największą p-value do momentu, w którym każda zmienna miała p-value mniejsze od ustalonego poziomu istotności 0.05.

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

aft_1 <- drop1(model_aft, test = "Chisq")
```

Tabela 1: Weryfikacja istotności zmiennych w modelu AFT - krok 1

Zmienna	p-value
age	0.2014
as.factor(sex)	0.0006
as.factor(ph.ecog)	0.0021
ph.karno	0.1332

Jak możemy zauważyć na tabeli 1 największą wartość p-value ma zmienna `age`, więc następny model budujemy właśnie bez niej.

```
model_aft_2 <- survreg(Surv(time, status) ~ as.factor(sex) + as.factor(ph.ecog)
                       + ph.karno, data = dane, dist = "weibull")

aft_2 <- drop1(model_aft_2, test = "Chisq")
```

Tabela 2: Weryfikacja istotności zmiennych w modelu AFT - krok 2

Zmienna	p-value
as.factor(sex)	0.0007
as.factor(ph.ecog)	0.0020
ph.karno	0.1756

Tym razem, na podstawie tabeli 2 widzimy, że największą wartość p-value miała zmienna `ph.karno`. Budujemy zatem nowey model bez tej zmiennej.

```
model_aft_3 <- survreg(Surv(time, status) ~ as.factor(sex) + as.factor(ph.ecog),
                       data = dane, dist = "weibull")

aft_3 <- drop1(model_aft_3, test = "Chisq")
```

Tabela 3: Weryfikacja istotności zmiennych w modelu AFT - krok 3

Zmienna	p-value
as.factor(sex)	0.0011
as.factor(ph.ecog)	0.0004

Widzimy, że w tabeli 3 zostały tylko dwie zmienne i każda z nim ma wartość p-value mniejszą od ustalonego poziomu istotności 0.05. Zatem wybranymi charakterystykami w modelu będą `sex` oraz `ph.ecog`. Zmienna ECOG ma mniejszą wartość p niż płeć. Sugeruje to, że kliniczna ocena sprawności pacjenta jest statystycznie jeszcze silniejszym sygnałem dla modelu niż różnice wynikające z płci, choć oba czynniki są kluczowe.

Podpunkt (b)

Do wyboru najlepszego modelu przyspieszonego czasu awarii wykorzystamy teraz kryterium informacyjne `AIC`.

```
model_aic <- step(model_aft, direction = "backward")
```

```
## Start:  AIC=2266.6
## Surv(time, status) ~ age + as.factor(sex) + as.factor(ph.ecog) +
##     ph.karno
##
##                       Df    AIC
## - age                  1 2266.2
## <none>                   2266.6
## - ph.karno             1 2266.8
## - as.factor(ph.ecog)   3 2275.2
## - as.factor(sex)       1 2276.4
##
## Step:  AIC=2266.23
## Surv(time, status) ~ as.factor(sex) + as.factor(ph.ecog) + ph.karno
##
##                       Df    AIC
## - ph.karno             1 2266.1
## <none>                   2266.2
## - as.factor(ph.ecog)   3 2275.0
## - as.factor(sex)       1 2275.8
##
## Step:  AIC=2266.07
## Surv(time, status) ~ as.factor(sex) + as.factor(ph.ecog)
##
##                       Df    AIC
## <none>                   2266.1
## - as.factor(sex)       1 2274.7
## - as.factor(ph.ecog)   3 2278.4
```

Krok 1:

- zaczynamy z `AIC` = 2266.6, model jest pełny
- funkcja widzi, że usunięcie zmiennej `age` obniży `AIC` do 2266.2, więc usuwamy zmienną `age`.

Krok 2:

- mamy model `sex` + `ph.ecog` + `ph.karno`
- funkcja widzi, że usunięcie `ph.karno` obnniży `AIC` do poziomu 2266.1

Krok 3:

- mamy model `sex` + `ph.ecog`
- Jeśli usuniemy `sex`, `AIC` podniesie się do 2274.7, a jeśli usuniemy `ph.ecog`, podniesie się do 2278.4.
- Zatem żadna redukcja nie poprawi już modelu.

Zatem zgodnie z kryterium `AIC`, najbardziej optymalnym modelem przyspieszonego czasu awarii jest ten zawierający zmienne `sex` oraz `ph.ecog`.

Podpunkt (c)

Korzystając z bayesowskiego kryterium informacyjnego (`BIC`), dokonamy wyboru najlepszego modelu przyspieszonego czasu awarii.

```
R <- sum(dane$status)
bic_aft <- step(model_aft, direction="backward", k = log(R))
```

```
## Start:  AIC=2291.35
## Surv(time, status) ~ age + as.factor(sex) + as.factor(ph.ecog) +
##     ph.karno
##
##                       Df    AIC
## - age                  1 2287.9
## - ph.karno             1 2288.5
## - as.factor(ph.ecog)   3 2290.7
## <none>                   2291.3
## - as.factor(sex)       1 2298.0
##
## Step:  AIC=2287.89
## Surv(time, status) ~ as.factor(sex) + as.factor(ph.ecog) + ph.karno
##
##                       Df    AIC
## - ph.karno             1 2284.6
## - as.factor(ph.ecog)   3 2287.4
## <none>                   2287.9
## - as.factor(sex)       1 2294.4
##
## Step:  AIC=2284.63
## Surv(time, status) ~ as.factor(sex) + as.factor(ph.ecog)
##
##                       Df    AIC
## <none>                   2284.6
## - as.factor(ph.ecog)   3 2287.7
## - as.factor(sex)       1 2290.2
```

Podobnie jak w podpunkcie (b), analizę rozpoczęto od pełnego modelu `AFT` uwzględniającego zmienne `age`, `sex`, `ph.ecog` oraz `ph.karno`. Następnie, w kolejnych krokach stopniowo usuwano zmienne w celu optymalizacji kryterium informacyjnego. Najpierw odrzucono `age`, a następnie `ph.karno`. Ostatecznie uzyskano model taki sam jak ten przedstawionym w podpunkcie (a). Skutkuje to taką samą interpretacją wyników. Funkcja `step` wyświetla wartości `AIC`, jednak dla parametru $k = log(R)$, gdzie R oznacza liczbę pełnych obserwacji, wyznaczana jest wartość `BIC`.

1.2 Selekcja cech w modelu Coxa metodą eliminacji krokowej

Przyjmiemy model proporcjonalnych hazardów Coxa. Jako zmienną zależną przyjęto zmienną `time`, natomiast do opisu charakterystyk pacjentów wykorzystano zmienne `age`, `sex`, `ph.ecog` oraz `ph.karno`.

Podpunkt (a)

Korzystając z metody eliminacji, w oparciu o test ilorazu wiarogodności, dokonano wyboru zmiennych do modelu przyspieszonego czasu awarii. Na początku zbudowano pełny model Coxa i w każdym kroku eliminowano zmienną z największą p-value do momentu, w którym każda zmienna miała p-value mniejsze od ustalonego poziomu istotności 0.05.

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

cox_1 <- drop1(model_cox, test = "Chisq")
```

Tabela 4: Weryfikacja istotności zmiennych w modelu Coxa - krok 1

Zmienna	p-value
age	0.1804
as.factor(sex)	0.0006
as.factor(ph.ecog)	0.0036
ph.karno	0.1886

Jak możemy zauważyć na tabeli 4 największą wartość p-value ma zmienna `ph.karno`, więc następny model budujemy właśnie bez niej.

```
model_cox2 <- coxph(Surv(time, status) ~ age + as.factor(sex) +
                     as.factor(ph.ecog), data = dane)

cox_2 <- drop1(model_cox2, test = "Chisq")
```

Tabela 5: Weryfikacja istotności zmiennych w modelu Coxa - krok 2

Zmienna	p-value
age	0.2343
as.factor(sex)	0.0010
as.factor(ph.ecog)	0.0010

Tym razem, na podstawie tabeli 5 widzimy, że największą wartość p-value miała zmienna `age`. Budujemy zatem nowey model bez tej zmiennej.

```
model_cox3 <- coxph(Surv(time, status) ~ as.factor(sex) + as.factor(ph.ecog),
                    data = dane)

cox_3 <- drop1(model_cox3, test = "Chisq")
```

Tabela 6: Weryfikacja istotności zmiennych w modelu Coxa - krok 3

Zmienna	p-value
as.factor(sex)	1e-03
as.factor(ph.ecog)	4e-04

Widzimy, że w tabeli 6 zostały tylko dwie zmienne i każda z nim ma wartość p-value mniejszą od ustalonego poziomu istotności 0.05. Zatem wybranymi charakterystykami w modelu będą `sex` oraz `ph.ecog`. Są to te same zmienne, jak w przypadku modelu `AFT`. Zmienną `ph.ecog` możemy uznać za najsilniejszy predyktor w modelu, ponieważ ma najniższe p-value. Sugeruje to, że kliniczna ocena tego, jak pacjent radzi sobie z codziennymi czynnościami, jest najlepszym wskaźnikiem tego, jakie ma szanse na przeżycie.

Podpunkt (b)

Do wyboru najlepszego modelu przyspieszonego czasu awarii wykorzystamy teraz kryterium informacyjne `AIC`.

```
aic_cox <- step(model_cox, direction="backward")
```

```
## Start:  AIC=1458.54
## Surv(time, status) ~ age + as.factor(sex) + as.factor(ph.ecog) +
##     ph.karno
##
##                       Df    AIC
## - ph.karno             1 1458.3
## - age                  1 1458.3
## <none>                   1458.5
## - as.factor(ph.ecog)   3 1466.0
## - as.factor(sex)       1 1468.2
##
## Step:  AIC=1458.27
## Surv(time, status) ~ age + as.factor(sex) + as.factor(ph.ecog)
##
##                       Df    AIC
## - age                  1 1457.7
## <none>                   1458.3
## - as.factor(sex)       1 1467.0
## - as.factor(ph.ecog)   3 1468.5
##
## Step:  AIC=1457.68
## Surv(time, status) ~ as.factor(sex) + as.factor(ph.ecog)
##
##                       Df    AIC
## <none>                   1457.7
## - as.factor(sex)       1 1466.5
## - as.factor(ph.ecog)   3 1470.1
```

Analizę rozpoczęto od pełnego modelu Coxa, obejmującego zmienne `age`, `sex`, `ph.ecog` oraz `ph.karno`. Następnie, w kolejnych krokach eliminacji, stopniowo usuwano zmienne w celu uzyskania lepszej wartości kryterium `AIC`. Najpierw odrzucono `ph.karno`, a następnie `age`. Po zakończeniu tego procesu uzyskano model identyczny z przedstawionym w punkcie (a), co pozwala na taką samą interpretację wyników.

Podpunkt (c)

Wykorzystując funkcję `step` dokonamy wyboru zmiennych do modelu Coxa, korzystając z kryterium `BIC`.

```
R <- sum(dane$status) # liczba obserwacji kompletnych
bic_cox <- step(model_cox, direction="backward", k = log(R))
```

```
## Start:  AIC=1477.1
## Surv(time, status) ~ age + as.factor(sex) + as.factor(ph.ecog) +
##     ph.karno
##
##                       Df    AIC
## - ph.karno             1 1473.7
## - age                  1 1473.8
## - as.factor(ph.ecog)   3 1475.3
## <none>                   1477.1
## - as.factor(sex)       1 1483.7
##
## Step:  AIC=1473.73
## Surv(time, status) ~ age + as.factor(sex) + as.factor(ph.ecog)
##
##                       Df    AIC
## - age                  1 1470.0
## <none>                   1473.7
## - as.factor(ph.ecog)   3 1474.7
## - as.factor(sex)       1 1479.4
##
## Step:  AIC=1470.05
## Surv(time, status) ~ as.factor(sex) + as.factor(ph.ecog)
##
##                       Df    AIC
## <none>                   1470.0
## - as.factor(ph.ecog)   3 1473.2
## - as.factor(sex)       1 1475.7
```

Analogicznie do punktu (b), procedurę rozpoczęto od pełnego modelu Coxa, obejmującego zmienne `age`, `sex`, `ph.ecog` oraz `ph.karno`. W kolejnych krokach, w celu poprawy wartości kryterium `BIC`, stopniowo eliminowano poszczególne zmienne, najpierw `ph.karno`, a następnie `age`. Po zakończeniu tej procedury uzyskano model identyczny z tym przedstawionym w punktach (a) i (b), co pozwala na taką samą interpretację wyników.

Warto zauważyć, że uzyskane wnioski są spójne ze wcześniejszymi obserwacjami. W modelach wybrano te zmienne, dla których nie było podstaw do odrzucenia hipotezy sugerującej ich istotny wpływ na analizowany czas przeżycia.