Biostatystyka — projekt końcowy

Jakub Zbrzezny

12 czerwca 2021

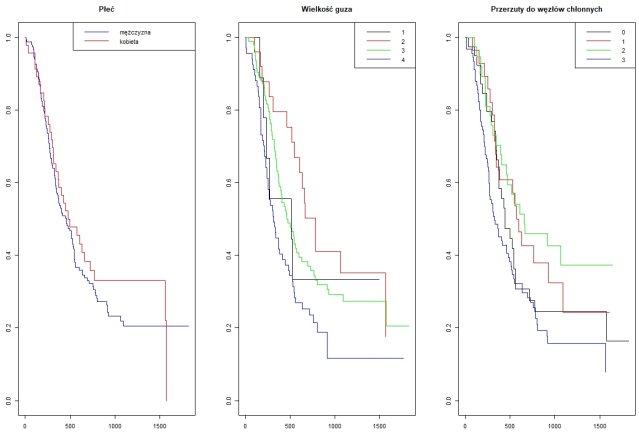
1 Wstęp

Udostępnione do analizy dane zawierają informacje dotyczące 195 chorych na raka części ustnej gardła włączonych do próby klinicznej porównującej czas przeżycia pacjentów leczonych radioterapią lub radioterapią i chemioterapią. Obok przedstawione są zmienne zawarte w zbiorze. Celem projektu jest porównanie funkcji przeżycia dla grup za pomocą odpowiednich testów oraz zbadanie wpływu zmiennych: `SEX`, `T_STAGE`, `N_STAGE`, następnie dopasowanie modelu Coxa oraz AFT, który jak najlepiej będzie opisywał wpływ poszczególnych zmiennych na czas przeżycia pacjentów.
W analizowanym zbiorze jest 195 obserwacji oraz nie ma braków danych. Mam do czynienia ze zdarzeniami jednego typu oraz nie nawracającymi.

`CASE`	identyfikator chorego
`INST`	identyfikator instytucji
`SEX`	płeć 1=mężczyzna, 2=kobieta
`TX`	leczenie 1=radioterapia, 2=radioterapia i chemioterapia
`GRADE`	stopień zróżnicowania nowotworu 1=wysoki, 2 = średni, 3=niski, 9=brak danych
`AGE`	wiek (w latach) w chwili diagnozy
`COND`	stopień sprawności chorego 1=bez ograniczeń, 2=ograniczony w pracy, 3=wymaga częściowej opieki, 4=wymaga całkowitej opieki (100% czasu w łóżku), 9=brak danych
`SITE`	lokalizacja guza 1=łuk podniebienny, 2=dół migdałkowy, 3=mięsień podniebienno-gardłowy, 4=nasada języka, 5=tylna ściana
`T_STAGE`	wielkość guza 1=2 cm lub mniej, 2 = 2-4 cm, 3=większy niż 4 cm, 4=masywny guz z naciekiem na okoliczne tkanki
`N_STAGE`	przerzuty do węzłów chłonnych 0=bez przerzutów, 1=jeden zajęty węzeł mniejszy niż 3 cm, ruchomy, 2=jeden zajęty węzeł większy niż 3 cm, ruchomy, 3=kilka zajętych i/lub nieruchomych węzłów
`STATUS`	wskaźnik zdarzenia 0=żyje, 1=zmarł
`TIME`	czas przeżycia w dniach od daty diagnozy

2 Porównania funkcji przeżycia dla grup

Najpierw porównałem funkcje przeżycia dla grup, z wykorzystaniem testu logrank. Sprawdziłem wpływ zmiennych: `AGE`, `T_STAGE`, `N_STAGE`.
Krzywe przeżycia prezentują się następująco:



W przypadku zmiennej określającej płeć, w pierwszej połowie czasu, krzywe przeżycia prawie się pokrywają, później prawdopodobieństwo przeżycia przez kobiety zaczyna wyraźnie przeważać nad prawdopodobieństwem przeżycia przez mężczyzn.
Jeśli chodzi o wielkość guza, to generalnie prawdopodobieństwo przeżycia jest największe dla wielkości guza 2-4

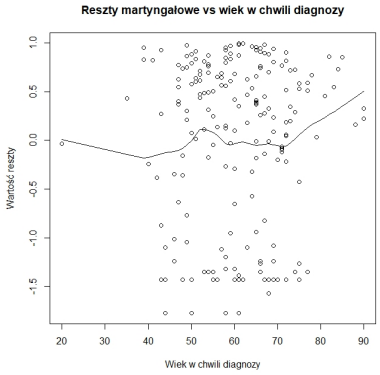
cm, później dla wielkości większej od 4 cm, potem dla wielkości 2 cm, a najmniejsze jest dla masywnego guza z naciekiem na okoliczne tkanki.
Natomiast w przypadku przerzutów do węzłów chłonnych, ogólnie prawdopodobieństwo przeżycia jest największe w przypadku jednego zajętego węzła większego niż 3 cm, ruchomego, później dla jednego zajętego węzła mniejszego niż 3 cm, ruchomego, potem dla pacjentów bez przerzutów, a najmniejsze jest dla kilku zajętych i/lub nieruchomych węzłów.
Przecięcia krzywych przeżycia generalnie są bardzo małe, prawie niezauważalne, a większe przecięcia są rzadko, zatem sprawdzę istotność statystyczną wpływów czynników na czas przeżycia za pomocą testu logrank. P-wartości dla przedstawionych 3 zmiennych, są w poniższej tabelce:

Zmienna	SEX	T_STAGE	N_STAGE
P-wartość	0.4	0.01	0.01

Zatem p-wartość dla zmiennej SEX jest większa od 0.05, a p-wartości dla zmiennych T_STAGE, N_STAGE, są mniejsze od 0.05. Stąd płeć nie ma statystycznie istotnego wpływu na czas przeżycia, ale wielkość guza oraz obecność przerzutów do węzłów chłonnych, ma wpływ na czas przeżycia.

3 Dopasowanie modelu Coxa

Przed przystąpieniem do modelowania, sprawdziłem postać funkcjonalną wszystkich zmiennych ciągłych objaśniających, aby je włączyć do modelu w odpowiedniej postaci. Jest nią tylko wiek (w latach) w chwili diagnozy. W tym celu, stworzyłem model pusty oraz skonstruowałem wykres reszt martyngałowych dla tej zmiennej.
Widać drobne odstępstwa od liniowości w środku oraz na prawym końcu, ale mimo ich, można włączyć zmienną AGE w postaci liniowej do modelu.



Analizę rozpocząłem od dopasowania modelu pełnego, którego formuła wygląda następująco:

```
Surv(TIME, STATUS) ~ as.factor(INST) + as.factor(SEX) + TX + GRADE + AGE + COND +
                     as.factor(SITE) + T_STAGE + as.factor(N_STAGE)
```

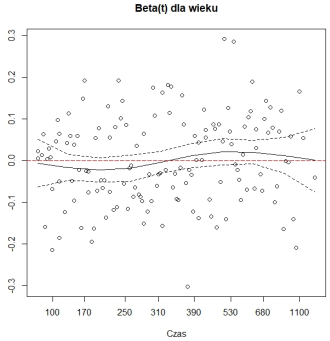
Następnie, przy pomocy testu Schoenfelda, sprawdziłem założenie proporcjonalnych hazardów dla tego modelu, otrzymując następujące p-wartości:

	as.factor(INST)	as.factor(SEX)	TX	GRADE	AGE	COND	as.factor(SITE)
P-wartość	0.178	0.828	0.221	0.900	0.059	0.547	0.403

T_STAGE	as.factor(N_STAGE)	GLOBAL
0.237	0.464	0.122

Widzimy, że dla każdej zmiennej, p-wartość jest większa od 0.05, zatem założenie proporcjonalnych hazardów jest spełnione. Jednakże, widać, że p-wartość dla zmiennej oznaczającej wiek w chwili diagnozy, jest widocznie odstająca od pozostałych wartości. Zatem chciałem sprawdzić liniowość współczynnika $\beta(t)$ dla AGE, za pomocą wykresu skalowanych reszt Schoenfelda.

Widzimy, że wartość układa się nieliniowo i nie pokrywa się z oszacowanym, z modelu współczynnikiem. Zatem postanowiłem skategoryzować zmienną AGE na 4 kategorie za pomocą kwantyli rzędu: 0.25, 0.5, 0.75. Kwantyle tych rzędów dla tej zmiennej, są równe: 52, 60, 68, odpowiednio. Zmienną skategoryzowaną oznaczyłem jako AGE2.



Formuła wygląda następująco:

```
Surv(TIME, STATUS) ~ as.factor(INST) + as.factor(SEX) + TX + GRADE + strata(AGE2) +
                     COND + as.factor(SITE) + T_STAGE + as.factor(N_STAGE)
```

Okazało się po przeprowadzeniu testu Schoenfelda, że wszystkie zmienne znów mają p-wartości większe od 0.05, zatem model spełnia założenie proporcjonalnych hazardów.

Kolejna moja próba dopasowania modelu opierała się na spostrzeżeniu, że pacjenci pomiędzy różnymi ośrodkami mogą się bardziej różnić niż pacjenci wewnątrz jednego ośrodka. Dlatego zdecydowałem się dodatkowo dokonać klasteryzacji na zmiennej `INST`, gdyż klasteryzacja daje odporny estymator wariancji oraz dzięki temu, że klasteryzacja uwzględnia możliwe zróżnicowania pomiędzy ośrodkami, można zmniejszyć wariancję i tym samym, błędy estymacji. Formuła modelu wygląda w ten sposób:

```
Surv(TIME, STATUS) ~ cluster(INST) + as.factor(SEX) + TX + GRADE + strata(AGE2) +
                     COND + as.factor(SITE) + T_STAGE + as.factor(N_STAGE)
```

Sprawdziłem wynik testu Schoenfelda i zauważyłem, że znów p-wartości dla wszystkich zmiennych są większe od 0.05, zatem założenie PH jest spełnione.

Później zdecydowałem się również przeprowadzić selekcję zmiennych strategią top-down, która zaczyna analizy od modelu pełnego, a następnie iteracyjnie wybiera model o najmniejszej funkcji kary (przyjąłem, że było to kryterium Akaike). Należy też zwrócić uwagę na to, że taki sposób wyboru zmiennych włączanych do modelu, nie uwzględnia spełniania przez nie założenia proporcjonalnych hazardów. Otrzymałem model o poniższej formule:

```
Surv(TIME, STATUS) ~ COND + T_STAGE + as.factor(N_STAGE)
```

W poniższej tabeli, przedstawiam wartości kryterium AIC dla kolejno dopasowywanych przeze mnie modeli oraz modelu otrzymanego w wyniku selekcji top-down.

	Model I	Model II	**Model III**	Top-down
Kryterium AIC	1325.844	936.9776	**928.796**	1310.127

Kierując się wartościami przedstawionymi w powyższej tabeli, za ostateczny model Coxa, uznałem model z kategoryzacją i warstwowaniem po zmiennej `AGE2` oraz klasteryzacją po zmiennej `INST`.

4 Dopasowanie modelu AFT

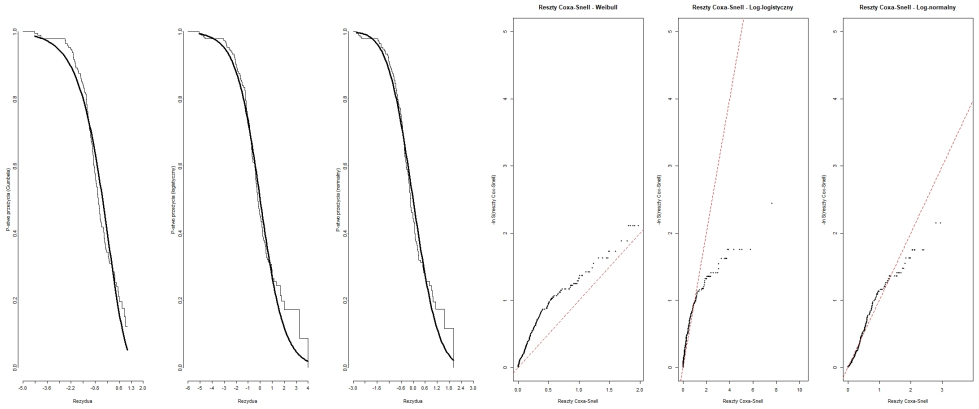
Modele AFT dopasowywałem z następującą formułą (dla wszystkich zmiennych):

```
Surv(TIME, STATUS) ~ INST + SEX + TX + GRADE + AGE + COND + SITE + T_STAGE + N_STAGE
```

Analizę rozpocząłem od dopasowania rozkładu uogólnionego F. Poniższa tabela przedstawia parametry wyestymowane z tego modelu, które pomogą w dopasowaniu mniej ogólnego rozkładu.

	est	L95%	U95%
mu	7.74e+00	5.10e+00	1.04e+01
sigma	1.79e-01	2.72e-02	1.19e+00
Q	-3.76e+00	-1.10e+01	3.52e+00
P	4.01e+01	9.19e-01	1.75e+03

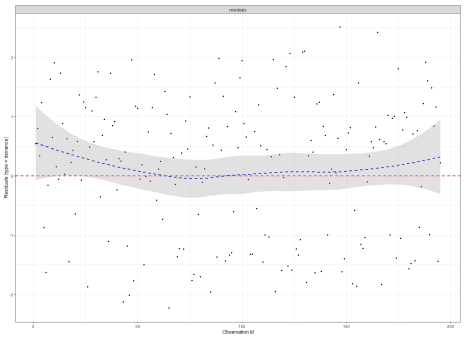
Można zauważyć na podstawie wyestymowanych parametrów, że rozkład uogólniony gamma nie będzie sugerowanym modelem, gdyż wyestymowany parametr `P` jest dużo większy od 0 (wynosi około 40), a wyestymowany parametr `Q` jest mniejszy od 0 (jest równy -3.76). W przypadku rozkładu Weibulla, log-logistycznego oraz log-normalnego, przeprowadziłem formalne testy ilorazu wiarogodności. Odrzucenie hipotezy zerowej przemawia za nieadekwatnością modelu. **P-wartości** dla modelu **Weibulla**, **log-logistycznego**, **log-normalnego**, wynoszą: **2.498289e-09**, **0.0002124637**, **6.544084e-05**, odpowiednio.
Zatem żaden z tych modeli nie będzie adekwatny, gdyż wszystkie p-wartości są znacznie mniejsze niż 0.05. Dla potwierdzenia, obliczam standaryzowane reszty. Dla modelu Weibulla powinny one odpowiadać (cenzurowanym) obserwacjom z rozkładu Gumbella, dla log-logistycznego (cenzurowanym) obserwacjom z rozkładu logistycznego, a dla log-normalnego (cenzurowanym) obserwacjom z rozkładu normalnego. Dla każdego modelu, tworzę wykres oszacowanej funkcji przeżycia dla reszt, a następnie nakładam na niego linię odpowiadającą rozkładowi adekwatnemu do wybranego modelu. Dodatkowo, przedstawię wykresy reszt Coxa-Snell dla tych modeli. Poniżej są przedstawione najpierw wykresy z funkcjami przeżycia, a potem wykresy reszt Coxa-Snell, kolejno dla modelu Weibulla, log-logistycznego, log-normalnego.



Zatem, rzeczywiście, w przypadku oszacowanych funkcji przeżycia, na żadnym z wykresów, krzywe teoretyczne, nie pokrywają się z wykresem oszacowanej funkcji przeżycia dla reszt.
Ponadto, także reszty Coxa-Snell potwierdzają przypuszczenie, że wszystkie te modele są nieadekwatne. Najlepiej prezentują się reszty dla log-logistycznego, gdyż tu najwięcej punktów jest bardzo blisko prostej $y = x$. Zauważmy, że dla tego modelu, p-wartość była największa spośród tych trzech. A najgorzej wyglądają dla Weibulla, ponieważ tu prawie wszystkie punkty wyraźnie odbiegają od prostej $y = x$. Potwierdza to również najmniejsza p-wartość spośród tych modeli AFT.

5 Wybór modelu ostatecznego, diagnostyka i interpretacja

Ostatecznie wybrałem model Coxa z kategoryzacją i warstwowaniem po zmiennej `AGE2` oraz klasteryzacją po zmiennej `INST`.
Sprawdzę dopasowanie modelu za pomocą wykresu reszt dewiancji.



Reszty są rozłożone w miarę symetrycznie wokół zera, jedynie na krańcach wykresu, są górne odchyłki od zera. Zatem model jest dość dobrze dopasowany.
Wyniki dla wybranego modelu, prezentują się następująco:

	coef	exp(coef)	se(coef)	robust se	z	p
as.factor(SEX)2	-0.20601	0.81382	0.20845	0.21499	-0.958	0.337941
TX	0.13268	1.14188	0.17544	0.16961	0.782	0.434075
GRADE	-0.10284	0.90227	0.12613	0.10316	-0.997	0.318827
COND	0.27370	1.31482	0.06919	0.07287	3.756	0.000173
as.factor(SITE)2	-0.26701	0.76566	0.21761	0.20329	-1.313	0.189035
as.factor(SITE)4	-0.17692	0.83785	0.22559	0.26840	-0.659	0.509800
T_STAGE	0.19807	1.21905	0.12852	0.08471	2.338	0.019372
as.factor(N_STAGE)1	-0.13850	0.87066	0.30346	0.19936	-0.695	0.487221
as.factor(N_STAGE)2	-0.05907	0.94264	0.32430	0.33611	-0.176	0.860483
as.factor(N_STAGE)3	0.41962	1.52138	0.23994	0.36028	1.165	0.244137

Stąd, z p-wartości dla poszczególnych zmiennych, widać, że istotnymi statystycznie zmiennymi są tylko zmienne `COND` oraz `T_STAGE`. Warto dodać, że te zmienne były wybrane po przeprowadzonej selekcji zmiennych top-down, przedstawionej w sekcji 3. Patrząc na eksponenty współczynników przy zmiennych, można zauważyć, że wzrost zmiennej stopnia sprawności chorego o jednostkę, wydłuża czas przeżycia o 31.5%. Przy wzroście zmiennej wielkości guza o jednostkę, czas przeżycia jest o 22% dłuższy niż wcześniej. Co więcej, czas przeżycia u kobiet jest o ponad 19% dłuższy niż u mężczyzn, ale mimo to, wzrost nie jest istotny statystycznie. Sposób leczenia, stopień zróżnicowania nowotworu, lokalizacja guza, przerzuty do węzłów chłonnych, mimo, że ich eksponenty są również wyraźnie różne od 1, nie mają statystycznie istotnego wpływu na czas przeżycia.