Projekt nr 5

Anna Herud

Cel projektu

Celem projektu jest analiza danych chorych na raka części ustnej gardła, porównując czas przeżycia pacjentów leczonych radioterapią lub radioterapią i chemioterapią. W projekcie przeanalizowany zostanie wpływ różnych zmiennych na czas przeżycia chorych oraz dopasowany zostanie model Coxa oraz model ATF.

Wstępna analiza zbioru danych

Dane wykorzystane w projekcie składają się z 195 obserwacji. Zmienna która nas interesuje, to zmienna `TIME`, czyli czas przeżycia pacjenta od chwili diagnozy. Najkrótszy czas przeżycia to 11 dni, a mediana szacowna estymatorem Kaplana - Meiera to 461 dni. Jest to jest najkrótszym, zaobserwowanym czasem zdarzenia, dla którego $S(t) < 50\%$.

Zmienna `STATUS` jest w naszym zbiorze danych indykatorem czasów cenzurowanych. Pacjentów, którzy nie zmarli w okresie prowadzonych obserwacji, uznajemy jako obserwacje cenzurowane. Około 27% danych, to obserwacje cenzurowane.

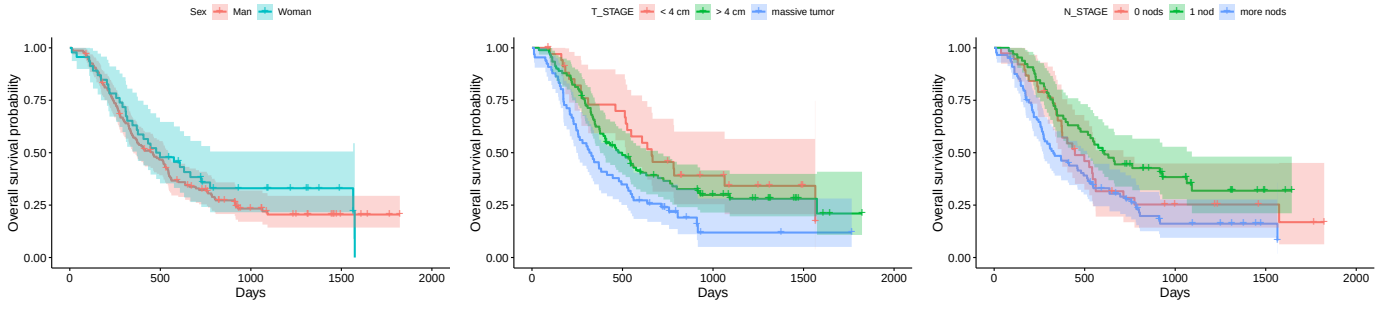
Zmienne, których wpływ na czas przeżycia pacjentów chorych na raka, to:

- `SEX`: płeć pacjenta (ok. 76% mężczyzn i ok. 24% kobiet)
- `GRADE`: stopień zróżnicowania nowotworu (1 obserwacja nie zawierająca danych, dla pozostałych: ok. 26% pacjentów o wysokim zróżnicowaniu nowotworu, ok. 56% o średnim zróżnicowaniu oraz ok. 18% procent o niskim stopniu zróżnicowania)
- `TX`: leczenie (ok. 51% pacjentów poddanych zostało radioterapii, a 49% pacjentów poddanych zostało radioterapii i chemioterapii)
- `AGE`: wiek w latach w chwili diagnozy
- `COND`: stopień sprawności chorego (2 obserwacje nie zwierają danych, dla pozostałych: ok. 74% pacjentów nie ma ograniczonej sprawności, ok. 22% ma ograniczoną sprawność w pracy, ok. 3% pacjentów wymaga częściowej opieki i 1 pacjent wymaga całkowitej opieki)
- `SITE`: lokalizacja guza (w przypadku ok. 33% pacjentów guz znajduje się na łuku podniebiennym, podobnie dla lokalizacji guza w dole migdałkowym, ok. 34% pacjentów posiada guza w nasadzie języka)
- `T_STAGE`: wielkość guza (ok. 5% pacjentów posiada guza o wielkości 2 cm lub mniej, ok. 13% o wielkości $2-4$ cm, ok. 48% o wielkości większej niż 4 cm, a ok. 23% pacjentów posiada masywnego guza z naciskiem na okoliczne tkanki)
- `N_STAGE`: przerzuty do węzłów chłonnych (20% pacjentów nie ma przerzutów, ok. 14% ma jeden zajęty węzeł mniejszy niż 3 cm - ruchomy, ok. 19% pacjentów ma jeden zajęty węzeł większy niż 3 cm - ruchomy, ok 47% pacjentów ma kilka zajętych węzłów)

Obserwacje dla których występują braku danych w zmiennej `GRADE` oraz `COND` zostaną usunięte.

Wpływ wybranych zmiennych na funkcję przeżycia

W pierwszej części zbadany zostanie wpływ zmiennych `SEX`, `N_STAGE` oraz `T_STAGE` na prawdopodobieństwo przeżycia pacjenta.



Wpływ płci na prawdopodobieństwo przeżycia

W pierwszej kolejności zbadana zostanie zmienna `SEX` składająca się z dwóch poziomów. Na wykresie dla zmiennej `SEX` została przedstawiona funkcja prawdopodobieństwa przeżycia dla każdej płci osobno. Na wykresie widać, że przedziały ufności krzywych przeżycia się pokrywają a same krzywe są dość blisko siebie. Krzywe przecinają się na początku i końcu wykresu, dlatego aby sprawdzić, czy prawdopodobieństwa przeżycia dla mężczyzn i dla kobiet się różnią istotnie zastosujemy test Renyiego, używając wag $W(t) = 1$. Testujemy następującą hipotezę:

$H_0 : \lambda_1(t) = \lambda_2(t)$

$H_1 : \lambda_1(t) \neq \lambda_2(t)$

gdzie $\lambda_1(t)$ to funkcja hazardu dla kobiet, a $\lambda_2(t)$ to funkcja hazardu dla mężczyzn.

Otrzymaliśmy duże p-value (0.27257), nie odrzucamy zatem hipotezy zerowej i zakładamy, że zmienna `SEX` nie ma wpływu na prawdopodobieństwo przeżycia.

Wpływ wielkości guza na prawdopodobieństwo przeżycia

Po narysowaniu funkcji prawdopodobieństwa przeżycia dla 4 poziomów, krzywa dla poziomu pierwszego (guz wielkości mniejszej niż 2 cm) przecina pozostałe krzywe i wygląda dość nieintuicyjnie. Zmienna `T_STAGE` dla poziomu 1 zawiera jedynie 5% obserwacji. Obserwacje z tego poziomu zostaną dołączone do poziomu drugiego, czyli guzów wielkości 2-4 cm.

Na wykresie dla zmiennej `T_STAGE` przedziały ufności funkcji przeżycia się pokrywają, a proste są zbliżone do siebie. Trudno jednak po samym wykresie stwierdzić, czy wielkość guza wpływa na prawdopodobieństwo przeżycia. W celu zweryfikowania hipotezy zostanie przeprowadzony test trendu. Badana hipoteza: $H_0 : \lambda_1(t) = \lambda_2(t) = \lambda_3(t)$

$H_1 : \lambda_1(t) \leq \lambda_3(t) \leq \lambda_4(t)$.

λ_i oznacza funkcję hazardu dla i-tej kategorii dla $i = 2, 3, 4$.

Małe p-value (0.00278) odrzuca hipotezę zerową. Zmienna `T_SATGE` ma wpływ na prawdopodobieństwo przeżycia. W szczególności, dla mniejszego guzu mamy większe prawdopodobieństwo przeżycia. Skoro wiemy, że zmienna `T_STAGE` wpływa na czas przeżycia, sprawdzimy również wpływ zmiennej `TX`, czyli rodzaju leczenia, na czas przeżycia w każdej z warstw zmiennej `T_STAGE`. Duże p-value (0.2337) wskazuje na to, że nie ma istotnych różnic pomiędzy czasem przeżycia ze względu na stosowanie różnych rodzai terapii dla różnych wielkości guza.

Wpływ przerzutów do węzłów chłonnych na prawdopodobieństwo przeżycia

Po narysowaniu funkcji przeżycia dla czterech poziomów zmiennej `N_STAGE` okazało się że krzywe dla poziomów 2 oraz 3 leżą bardzo blisko siebie. Obie z tych kategorii zawierają również mało obserwacji, zostaną one zatem połączone. Według wykresu wydaje się, że ilość węzłów z przerzutami ma wpływ na prawdopodobieństwo przeżycia, w szczególności pacjenci z 1 węzłem z przerzutami mają większe prawdopodobieństwo przeżycia niż pozostali pacjenci. W celu potwierdzenia tej intuicji przeprowadzony zostanie test log-rank:

$H_0 : \lambda_1(t) = \cdots = \lambda_k(t)$

$H_1 : \lambda_i(t) \neq \lambda_j(t)$

dla pewnych i, $j \in \{0, 2, 3\}$.

Małe p-value (0.00636)odrzuca hipotezę zerową o równości prawdopodobieństw przeżycia dla każdego z poziomu zmiennej `N_STAGE`. Podobnie jak w przypadku poprzedniej zmiennej, sprawdzone zostało, czy po warstwowaniu można zauważyć różnice we wpływie wyboru terapii na prawdopodobieństwo przeżycia dla różnych liczb węzłów z przerzutami. Ponownie wynik testu wskazuje, że nie ma istotnych różnic pomiędzy wpływem rodzaju zastosowanej terapii na czas przeżycia w różnych grupach pacjentów o różnej ilości węzłów chłonnych z przerzutami.

Model Coxa proporcjonalnych hazardów

W celu dalszej analizy wpływu zmiennych na czas przeżycia pacjentów, dopasowany zostanie model Coxa proporcjonalnych hazardów.

Dobór zmiennych

Zanim dopasowany zostanie pełny model, ze wszystkimi zmiennymi objaśniającymi, w zmiennej `COND` usunięty zostanie jeden poziom. Dla `COND = 4` występuje jedynie jedna obserwacja. Klasyfikujemy ją do klasy 3, osoby wymagające częściowej opieki i pełnej opieki będą teraz tak samo oznaczane.

Wynik `summary` dla pełnego modelu zwraca informację o tym, że większość zmiennych nie jest istotna statystycznie, według testu Walda. Selekcji zmiennych dokonamy za pomocą kryterium AIC.

Dopasowany model zawiera następujące zmienne: `SEX`, `COND` i `T_STAGE`.

Według testu Walda mamy zmienne nieistotne statystycznie, lecz na razie pozostawimy je w modelu.

Diagnostyka

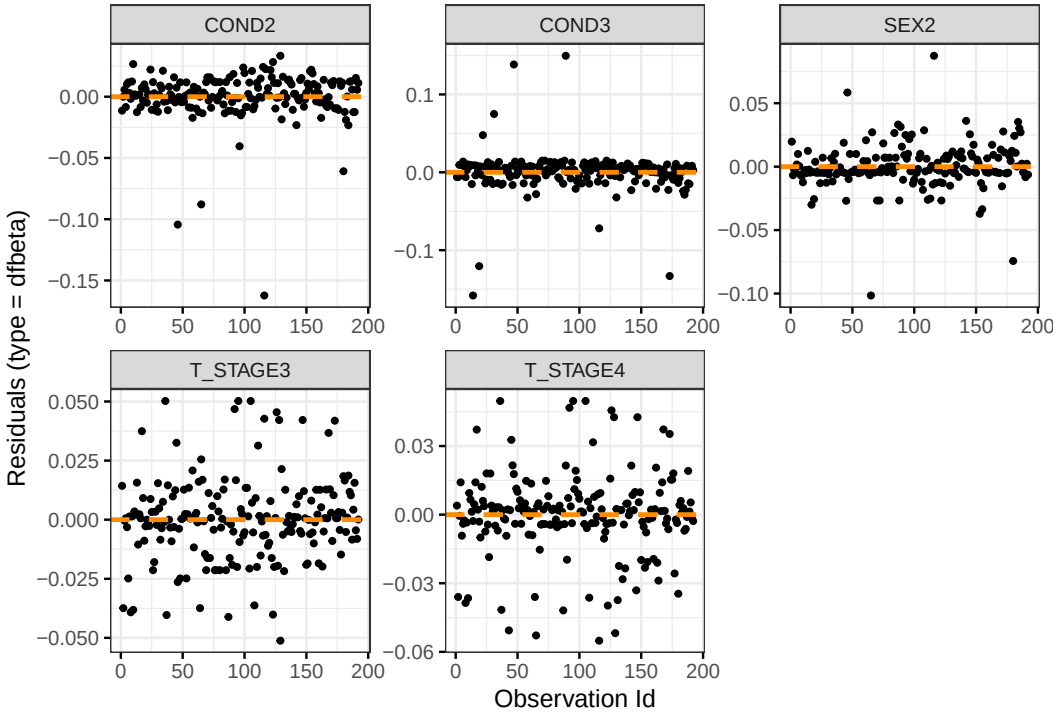
Przejdźmy teraz do diagnostyki naszego modelu. Kluczowym założeniem naszego modelu jest proporcjonalność hazardów. Pierwszym narzędziem, użytym w celu sprawdzenia tego założenia, będzie wykres residuów typu deviance.

W modelu, w którym założenie o proporcjonalnych hazardach jest spełnione wykres funkcji jest prostą zbliżoną do prostej na poziomie zera. Na wykresie dla naszych danych prosta odchyla się nieznacznie od zera. Jej odchylenie lub kształt może sugerować uwzględnienie pewnej zależności od czasu w modelu. W naszym przypadku wszystkie wybrane zmienne są kategoryczne, dlatego nie będziemy analizować wykresów diagnostycznych pod tym kątem. Z tego samego powodu pominięta zostanie diagnostyka za pomocą reszt Schoenfelda, które badają niezależność residuów (a zatem i zmiennych ciągłych) od czasu.

Przeprowadzony zostanie formalny test Schoenfelda, badający spełnienie założenia o proporcjonalności hazardów dla każdej ze zmiennych w modelu.

```
##           chisq df     p
## SEX      0.0126  1 0.911
## COND     9.0332  2 0.011
## T_STAGE  3.8241  2 0.148
## GLOBAL  11.6894  5 0.039
```

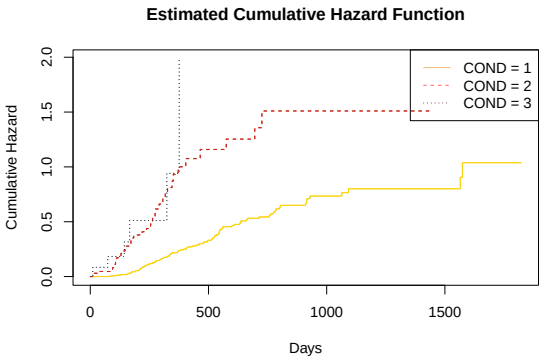
Wynik p-value testu mówi nam o tym, że założenie o proporcjonalności hazardu nie jest spełnione. Główny problem występuje w przypadku zmiennej `COND`. Zanim sprawdzone zostaną rozwiązania problemu przeanalizowany zostanie wykres reszty typu dfbeta w celu detekcji ewentualnych obserwacji wpływowych:



Wykres nie wydaje się wskazywać na obserwacje, które jednoznacznie można zakwalifikować jako wpływowe.

W celu identyfikacji obserwacji odstających użyty został się wykresem rezyduów typu deviance. Residua są rozmieszczone były równomiernie wokół zera, żadna z obserwacji wyraźnie nie odstaje, zakładamy zatem, że nie mamy obserwacji, którą należałoby usunąć jako odstającą.

Problem niespełnienia założeń o proporcjonalnych hazardach spróbujemy rozwiązać za pomocą warstwowania względem zmiennej `COND`, gdyż ona według wyniku testu Schoenfelda miała najgorszy wynik. Poniżej znajduje się wykres przedstawiający funkcję skumulowanego hazardu dla wartości każdego z poziomów `COND` (model proporcjonalnych hazardów, to również model proporcjonalnych skumulowanych hazardów). Krzywe się przecinają, wykres również wskazuje na niespełnienie założeń proporcjonalnych hazardów w przypadku zmiennej `COND`.



Spróbujemy zatem dopasować model z tymi samymi zmiennymi, jednak dokonując warstwowania według zmiennej `COND`.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
SEX2	-0.3560255	0.7004548	0.2127088	-1.6737697	0.0941759
T_STAGE3	-0.0447717	0.9562157	0.2565739	-0.1744983	0.8614738
T_STAGE4	0.5534027	1.7391607	0.2601873	2.1269396	0.0334251

Według testu Walda obie zmienne `T_STAGE` oraz `SEX` po przeprowadzeniu warstwowania są istotne. Sprawdzimy czy dla tak powstałego modelu założenia proporcjonalnych hazardów są spełnione, ponownie za pomocą testu Schoenfelda:

```
##          chisq df    p
## SEX     0.279  1 0.60
## T_STAGE 1.857  2 0.40
## GLOBAL  2.110  3 0.55
```

Duże wartości p-value nie odrzuca nam hipotezy zerowej o spełnieniu założenia proporcjonalności hazardów dla każdej ze zmiennych.

Dla nowo powstałego modelu również przeprowadzaona została diagnostyka mająca na celu wyłapanie potencjalnych obserwacji odstających lub wpływowych, lecz takie nie zostały znalezione.

Ostateczna postać modelu to:

$$\lambda(t) = \lambda_i(t) \cdot \exp(-0.356 \cdot X_{sex2} - 0.045 \cdot X_{t_stage3} + 0.553 \cdot X_{t_stage4})$$

dla $i = 1, 2, 3$, gdzie $\lambda_i(t)$ to hazard bazowy dla i-tego poziomu zmiennej `COND`.

Model ATF

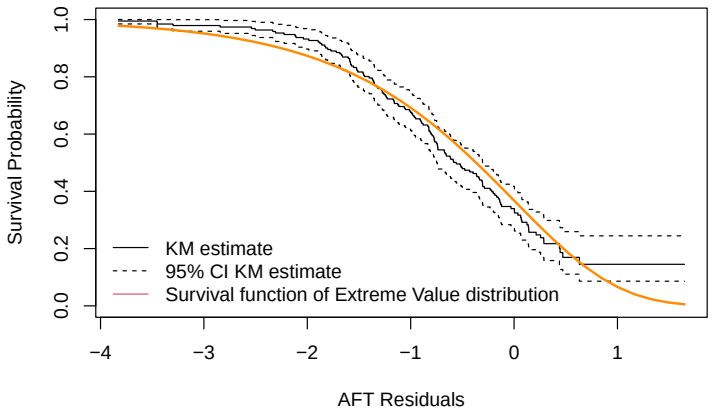
W ostatnim etapie analizy do danych zostanie dopasowany model AFT (ang. accelerated failure time), czyli model przyspieszonego czasu do niepowodzenia.

Dopasowanie modelu

w pierwszej kolejności do danych został dopasowany rozład uogólniony gamma w celu sprawdzenia, czy któryś rozkład z tej rodizny rozkładów będzie odpowiedni dla naszych dnaych. Wyboru zmiennych dokonujemy za pomocą kryterium AIC. Wybrane zmienne to: `T_STAGE`, `SEX` oraz `COND`, czyli te same zmienne co w przypadku modelu Coxa.

Sprawdzamy przedział ufności dla parametru Q aby sprawdzić czy zawiera on 1 (rozkład Weibulla) lub 0 (rozkład log normalny). Otrzymany przedział ufności na poziomie 5% to $(-0.8233599, 0.2036972)$. Przedział nie zawiera ani 1 ani 0. Dla pewności dopasowany został osobno również rozkład Wiebulla, jako że jest to najbardziej popularny rozkład dla modeli AFT. Jednak wykresy diagnostyczne potwierdzają złe dopasowanie rozkładu do do danych.

Poniżej wykres przedstawiający estymator Kaplana - Meiera obliczonego na podstawie rezyduów oraz dopasowany rozkład Weibulla.

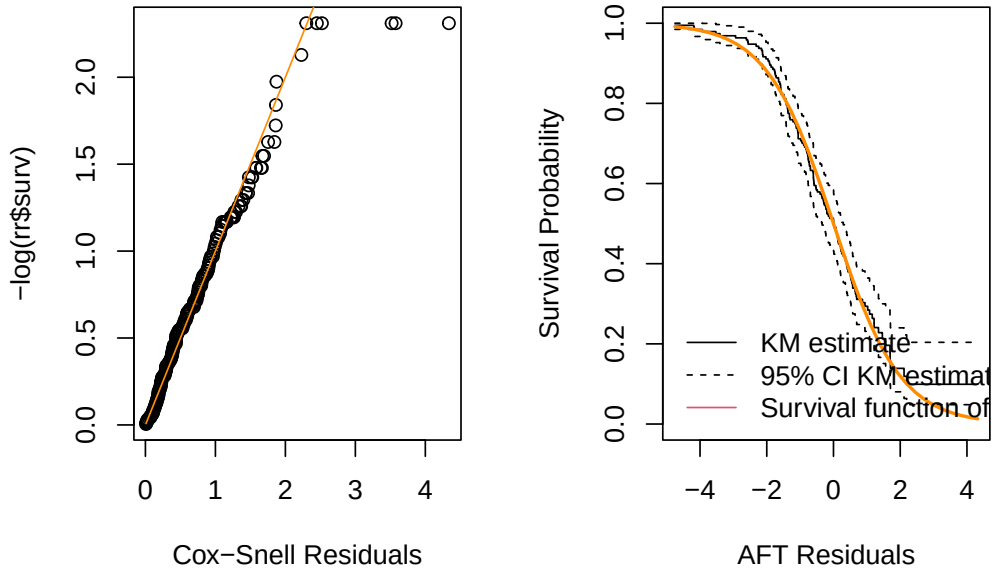


Wykres sugeruje, że rozkład Weibulla nie jest dobrym wyborem dla naszych danych. Czerwona krzywa odpowiadająca rozkładowi nie mieści się w przedziałach ufności estymatora KM.

Model AFT dla rozkładu log - logistic

Spróbujmy dopasować model dla rozkładu log-logistic. Selekcji zmiennych ponownie dokonujemy za pomocą kryterium AIC. Wybrane zmienne to: `COND`, `T_STAGE` oraz `N_STAGE`.

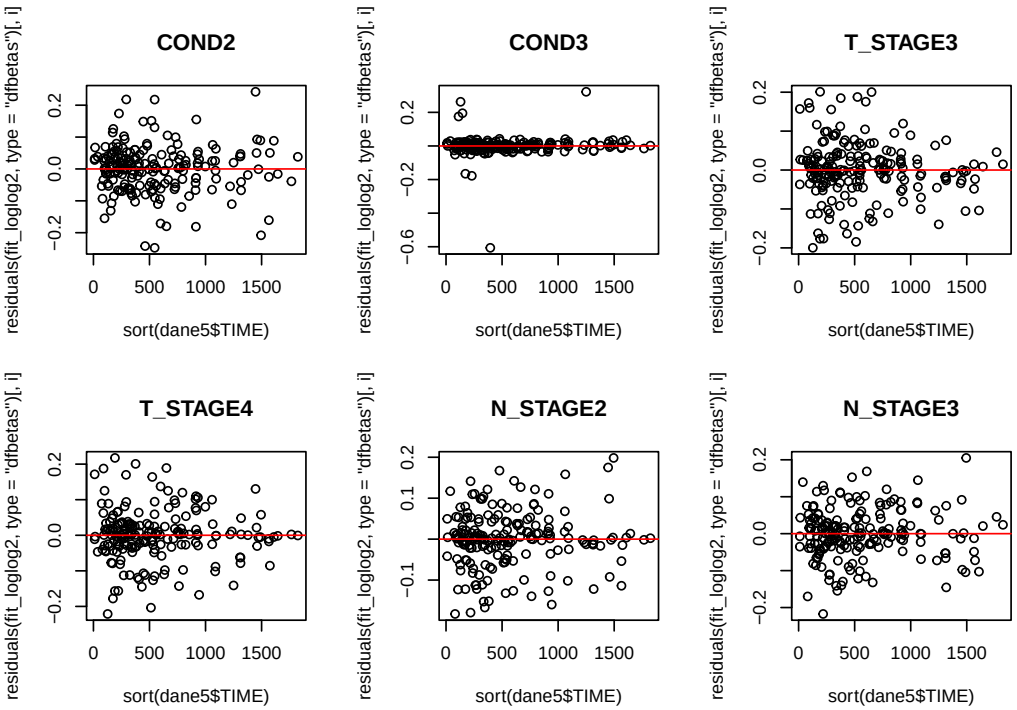
Na poniższym wykresie po lewej reszty Coxa Snella nie wskazują na większe problemy z dopasowaniem. W górnej części wykresu widać jedynie parę obserwacji odstających od prostej. Po prawej ponownie wykres estymatora KM dla residuów z dopasowanych rozkładem log-logistik. Wykres ten wskazuje na dużo lepsze dopasowanie niż w przypadku rozkładu Weibulla.



Obserwacje odstajace i wpływowe

Przejdziemy teraz do analizy wykresu residuów typu deviance w celu sprawdzenia czy występują jakieś obserwacje odstające. Wykres nie wykazał żadnych odstępstw od normy, rezydua były rozłożone równomiernie względem prostej $y = 0$.

Analizujemy również wykresy dla reszt typu dfbeta, w celu detekcji obserwacji wpływowych:



Dla zmiennej COND, dla poziomu 3 jedna obserwacja wydaje się wpływowa, o wielkości residuum ok -0.6. Zostanie ona usunięta ze zbioru danych. Dla modelu dopasowanego na pomniejszonym zbiorze wykresy diagnostycznie nie wskazują na obserwacje wpływowe.

Ostateczna postać modelu:

$$log(T) = 6.73 - 0.98 \cdot X_{cond2} - 1.07 \cdot X_{cond3} - 0.05 \cdot X_{t_stage3} - 0.52 \cdot X_{t_stage4} + 0.06 \cdot X_{n_stage2} - 0.29 \cdot X_{n_stage4} + 0.52\varepsilon$$

Podsumowanie

Przeprowadzona powyżej analiza danych pacjentów cierpiących na raka części ustnej gardła pozwoliła na wysnucie następujących wniosków:

- Zmienna określająca wielkość guza (`T_STAGE`) to zmienna, która we wszystkich trzech punktach analizy wykazała wpływ na prawdopodobieństwo przeżycia pacjenta. Wynik ten jest intuicyjny, wielkość guza świadczy o powadze stanu pacjenta.
- Pozostałe zmienne, które wykazały wpływ na prawdopodobieństwo przeżycia, na przynajmniej jednym z etapów analizy to: liczba węzłów z przerzutami (`N_STAGE`), sprawność pacjenta (`COND`) oraz płeć.
- Powstały model Coxa pozwala nam na następująca interpretacje wpływu płci oraz wielkości guza (dla każdego stopnia sprawności pacjenta osobno):
  - dla płci żeńskiej hazard jest 0.7 razy niższy niż dla płci męskiej ($HR_{sex2} = exp(-0.356)$)
  - dla guzów większych niż 3 cm hazard jest 0.95 niższy niż dla pozostałych grup zmiennej ($HR_{t_stage3} = exp(-0.045)$)
  - dla masywnych guzów z naciekiem na okoliczne tkanki hazard jest 1.74 razy wyższy niż w pozostałych grupach ($HR_{t_stage4} = exp(0.553)$).
- Powstały model AFT pozwala nam na następującą interpretację wpływu zmiennych określających sprawność pacjenta, wielkość guza oraz liczbę węzłów z przerzutami:
  - Czas przeżycia pacjenta skraca się dla osób o ograniczonej sprawności w pracy oraz dla osób wymagających opieki, przy czym skraca się bardziej dla osób w drugim przypadku.
  - Czas przeżycia pacjenta skraca się w zależności od wielkości guza, najbardziej dla osób o masywnych guzach z naciskiem na okoliczne tkanki (pokrywa się to z wynikami z modelu Coxa).
  - Czas przeżycia wydłuża się dla dla osób z jednym węzłem z przerzutami w porównaniu do innych kategorii, zaś wydłuża się dla pacjentów z kilkoma węzłami z przerzutami.