1. **Opis zbioru danych**

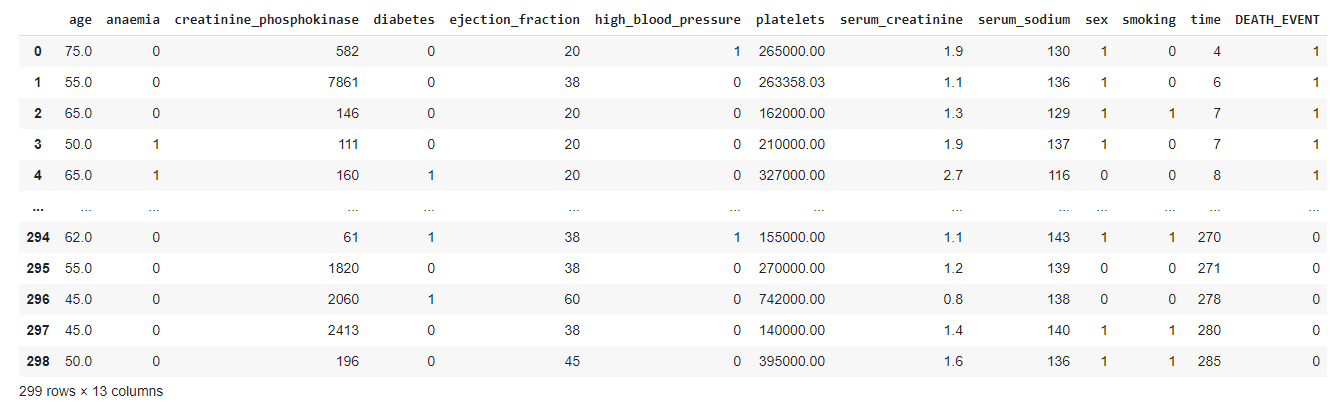
Zbiór danych pochodzi ze strony <https://www.kaggle.com/andrewmvd/heart-failure-clinical-data> i przedstawia obserwacje pacjentów cierpiących na choroby sercowo-naczyniowe. Choroby te są najczęstszą globalną przyczyną śmierci, szacuje się, że około 17.9 miliona ludzi umiera każdego dnia, co stanowi około 31% wszystkich śmierci w skali całej populacji świata. Dla każdej obserwacji w zbiorze danych jest 13 zmiennych, opisanych poniżej:

|  |  |
| --- | --- |
| **Zmienna** | **Opis** |
| Age | Wiek pacjenta (w latach). |
| Anaemia | Spadek erytrocytów lub hemoglobiny (0 – nie, 1 – tak). |
| creatinine\_phosphokinase | Poziom enzymu CPK we krwi (mcg/L). |
| Diabetes | Pacjent cierpi na cukrzyce (0 – nie, 1 – tak). |
| ejection\_fraction | Procent krwi opuszczający serce przy każdym uderzeniu. |
| high\_blood\_pressure | Pacjent ma zbyt wysokie ciśnienie krwi (0 – nie, 1- tak). |
| Platelets | Płytki krwi (w tysiącach na mL). |
| serum\_creatinine | Poziom kreatyniny w surowicy krwi (mg/dL). |
| serum\_sodium | Poziom sodu we krwi (mEq/L). |
| Sex | Płeć (kobieta – 0, mężczyzna – 1). |
| Smoking | Pacjent pali (0 – nie, 1 – tak). |
| Time | Czas śmierci lub końca obserwacji pacjenta (w dniach od rozpoczęcia obserwacji). |
| DEATH\_EVENT | Pacjent zmarł w czasie obserwacji (0 – nie, 1- tak). |

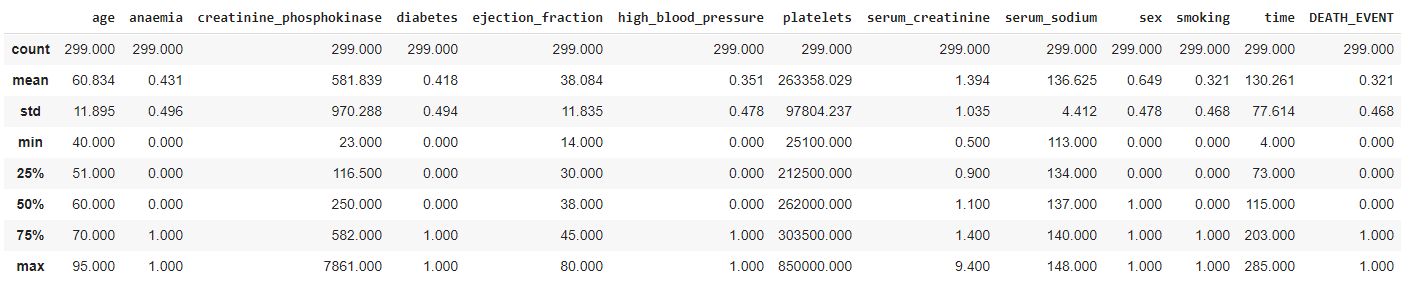
Zmienna DEATH\_EVENT będzie służyć do określenia realizacji badanego wydarzenia (śmierci pacjenta) w czasie time (w przypadku DEATH\_EVENT = 1) lub do cenzurowania obserwacji w czasie time (w przypadku DEATH\_EVENT = 0). Pozostałe zmienne będą zmiennymi objaśniającymi – rozkład wybranych (najistotniejszych) zmiennych objaśnianych zostanie dokładniej opisany w dalszej części pracy.

1. **Eksploracja danych**

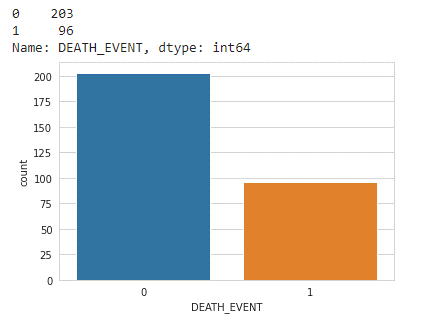
Pierwsze i ostatnie obserwacji zbioru danych prezentują się następująco:



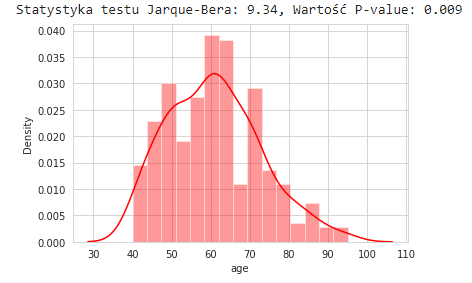
Statystyki opisowe wszystkich zmiennych:



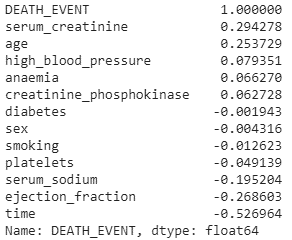
oraz rozkład zmiennej binarnej DEATH\_EVENT:



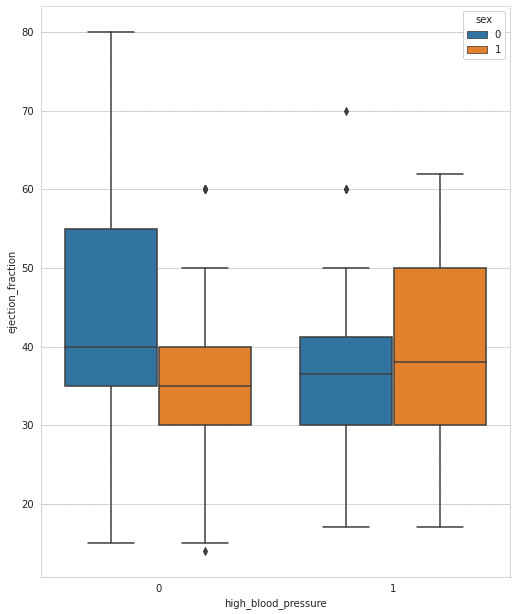
Zbiór składa się z 299 obserwacji, w tym 96 obserwacji śmierci pacjenta oraz 203 obserwacji cenzurowanych. Średnia wieku pacjentów to 60,8 lat, ze średnim odchyleniem równym 11,9 lat. Mediana wieku pacjentów wynosi 60 lat, najmłodszy obserwowany pacjent miał 40 lat, a najstarszy 95. Rozkład wieku pacjentów nie jest rozkładem normalnym, co zostało zbadane testem Jarque-Bera. Rozkład ten charakteryzuje się prawostronną asymetrią.



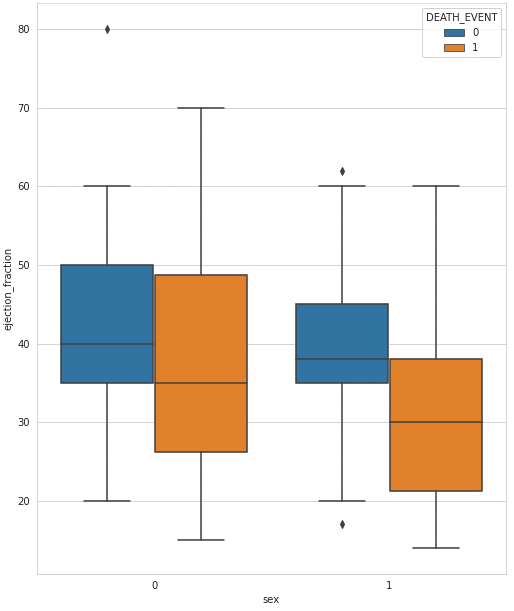
Następnie zbadana została korelacja zmiennych objaśniających ze zmienną DEATH\_EVENT, aby zlokalizować prawdopodobne predyktory zmienne objaśnianej.



Wartości współczynniku korelacji Pearsona wskazują na to, że największą moc predykcyjną mogą mieć zmienne wiek, kreatynina we krwi, frakcja wyrzutowa, oraz czas końca obserwacji (który z natury będzie brany pod uwagę w przypadku analizy przeżycia). Należy jednak wziąć pod uwagę, że są to jedynie zmienne podejrzane o wpływ, ponieważ korelacja niekoniecznie oznacza przyczynowość.



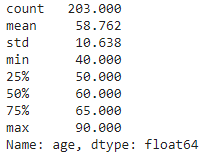
Przedstawiony powyżej wykres pudełkowy opisuje rozkład frakcji wyrzutowej w zależności od płci oraz tego czy pacjent ma zbyt wysokie ciśnienie krwi. Frakcja wyrzutowa zarówno wśród mężczyzn jak i wśród kobiet ma zbliżoną medianę oscylującą około 35-40 punktów procentowych. Rozkład frakcji wyrzutowej jest jednak bardzo rozciągnięty jedynie dla mężczyzn o niepodwyższonym ciśnieniu krwi, występuje zatem wiele obserwacji odstających – które potencjalnie znacznie częściej zakończyły badanie przed śmiercią pacjenta.

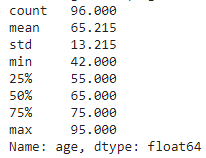


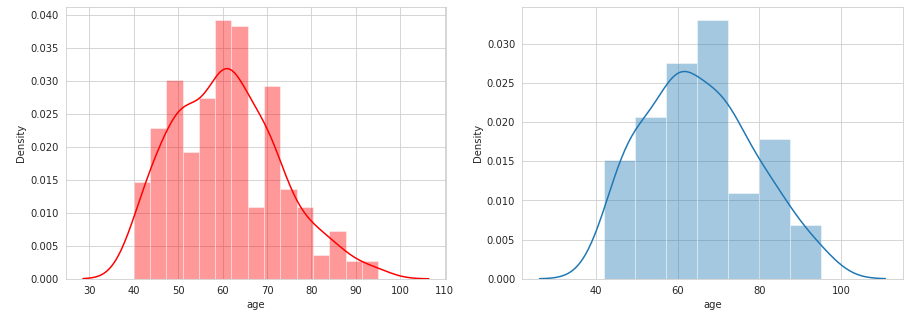
Kolejny wykres pudełkowy także opisuje frakcje wyrzutową, tylko w poróżnieniu poprzez płeć oraz to czy pacjent zmarł czy nie przed końcem obserwacji. Zarówno kobiety jak i mężczyźni, które umarły przed końcem obserwacji, miały mniejszą frakcję wyrzutową o około 5 i 8 punktów procentowych oraz bardziej rozciągnięty względem mediany rozkład z większą liczbą obserwacji odstających, od pacjentów którzy przeżyli do końca trwania obserwacji.

Opisane poniżej statystyki opisowe dotyczą zmiennej wiek, w podziale na pacjentów, którzy zmarli przed końcem obserwacji oraz tych, którzy przeżyli. Pacjenci, którzy zmarli mieli średnio 65 lat (ze zbliżoną medianą) z odchyleniem standardowym około 13,2 lat, podczas gdy pacjenci, którzy przeżyli do końca obserwacji mieli średnio około 58,7 lat (mediana 60 lat) z odchyleniem standardowym około 10 lat. Wiek nie wydaje się silnym predyktorem zmiennej objaśnianej, jednak zdaje się zmniejszać szansę na przeżycie do pewnego stopnia.

Poniżej kolejno statystyki opisowej oraz histogramy rozkładów pacjentów, którzy przeżyli oraz zmarłych.







1. **Estymator Kaplana-Meiera**

Za pomocą estymatora Kaplana-Meira, korzystając ze wzroru:

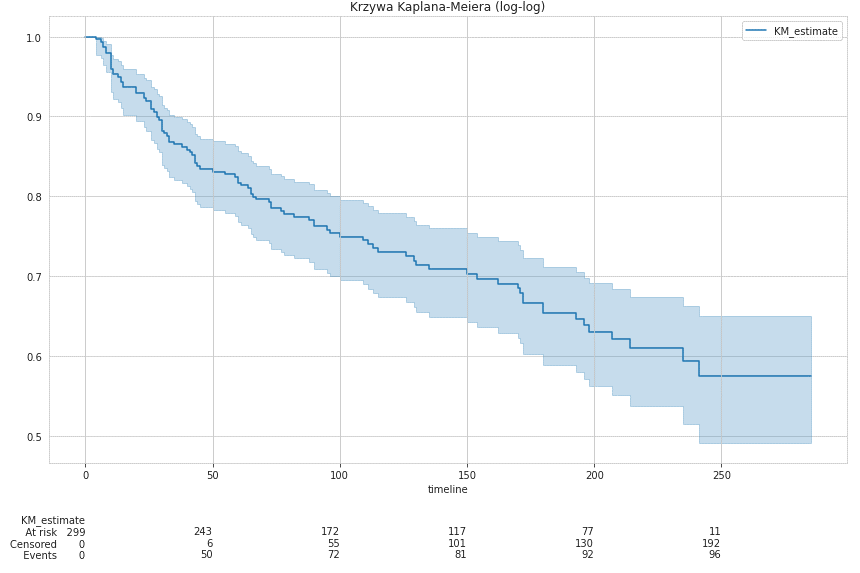
,

gdzie

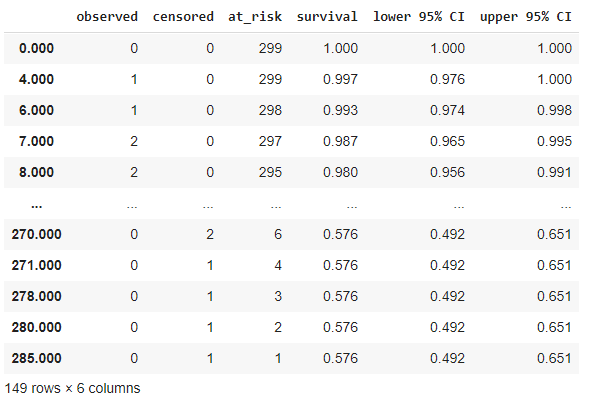
dla ,

oznacza estymator prawdopodobieństwa warunkowego w przedziale , pod warunkiem, że jednostka przetrwała wszystkie poprzednie przedziały, to ilość obserwacji, które dotrwały do momentu , a to liczba zdarzeń w tym momencie.

Estymator Kaplana-Meira jest estymatorem największej wiarygodności, asymptotycznie nieobciążonym, zgodnym i o rozkładzie asymptotycznie normalnym.

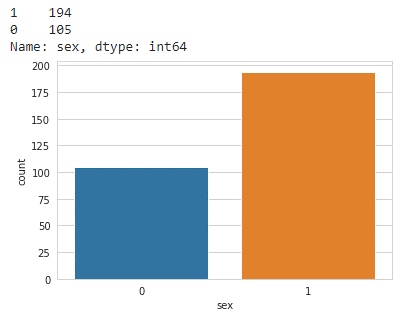


Powyższa krzywa oszacowana została dla zmiennej czasu obserwacji śmierci pacjenta bądź cenzurowania, oraz samego zdarzenia, wraz z przedziałami ufności 95% liczonymi metodą log-log. Z krzywej wynika, że prawdopodobieństwo przeżycia znacznie malej wraz z czasem obserwacji jednostki. Na początku prawdopodobieństwo maleje najszybciej, potem przy wartości około 0.85 zwalnia, ale wciąż dosyć szybko maleje z podobną dynamiką aż do około 200 dnia obserwacji pacjenta, kiedy jego szansa na przeżycie maleje o około 4-5 punktów procentowych prze kolejne 85 dni.

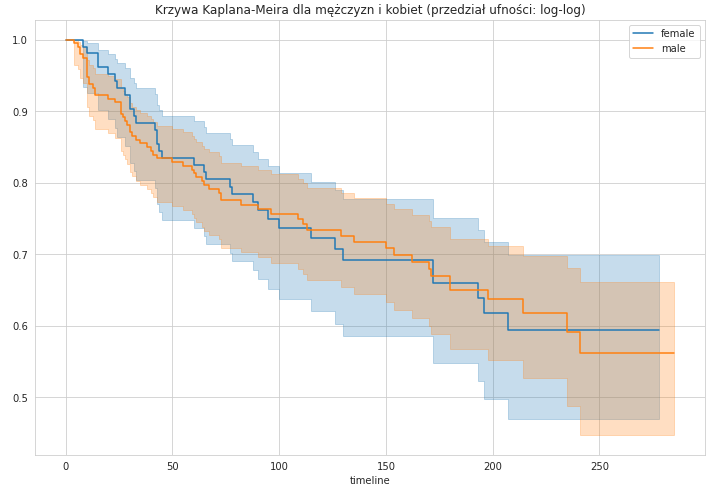


1. **Różnicowanie krzywej Kaplana-Meiera**

Następnym krokiem analizy jest różnicowanie krzywej poprzez zmienne objaśniające. Pierwszą taką zmienną jest płeć.

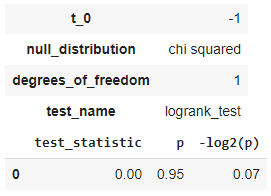


W zbiorze danych jest znacznie więcej obserwacji pacjentów płci męskiej – jest i 194, a obserwacji kobiet jest 105. Bazując na obliczonych wcześniej współczynnikach korelacji Pearsona oraz przeprowadzonej eksploracji danych, nie ma podejrzeń co do istotnego wpływu tej zmiennej na prawdopodobieństwo przeżycia, jednak jest to często różnicująca zmienna w medycynie.



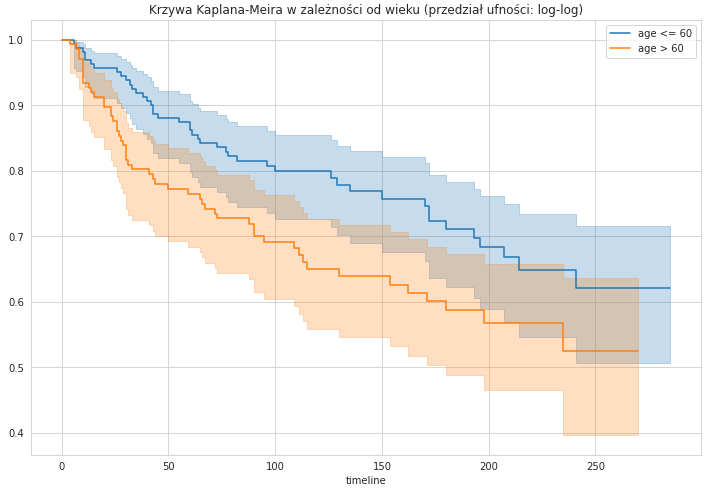
Powyższy wykres z krzywymi potwierdza to przypuszczenie. Następnie wykonany zostanie test Log-rank na równość tych krzywych. Ten wariant testu został wybrany ze względu na równe wagi dla całej funkcji trwania – której wartości w analizowanym przykładzie wydają się być podobne dla obu grup przez cały czas trwania obserwacji.

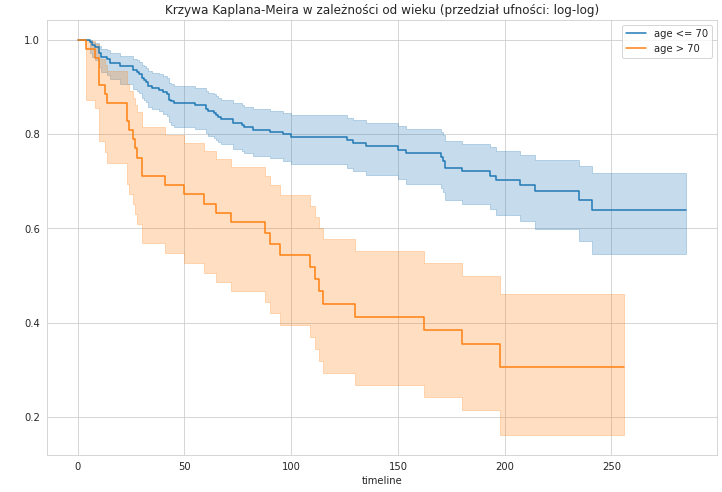
Hipotezy testów na równość funkcji trwania dwóch grup opisane są jako:



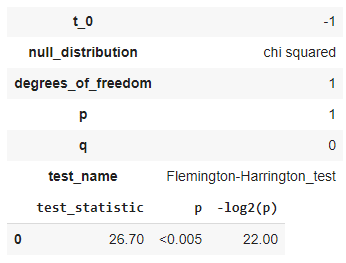
Test wykazał bardzo niską statystkę testową oraz wysokie p-value, co nie daje podstaw do odrzucenia hipotezy zerowej, mówiącej o równości funkcji przeżycia w danych grupach (ze względu na płeć).

Drugą badaną zmienną objaśniającą jest wiek, której rozkład został już opisany na stronie 3 i 6. Funkcja przeżycia została zróżnicowana zależnie od tego czy pacjent miał mniej niż 60 lat (lub równo 60 lat), czy też więcej.

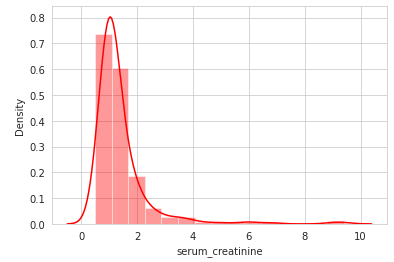


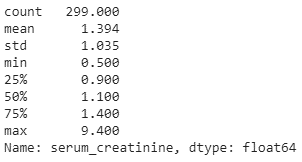
Wizualna analiza wskazuje na to, że istnieje istotna różnica pomiędzy dwoma wariantami. Następnie zwiększono próg 60 lat (który jest średnią wieku badanych pacjentów) do 70 lat. 

Znacznie zwiększyło to wizualne różnice pomiędzy krzywimi trwania – osoby w wieku powyżej 70 lat mają znacznie niższe prawdopodbieństwo przeżycia. Ze względu na to, że wartości funkcji trwania spada zdecydowanie szybciej na początku wśród starszych pacjentów, użyty zostanie test dający większe wagi na początku funkcji. Jest to test Fleminga-Harringotna ze współczynnikami p=1 i q=0.

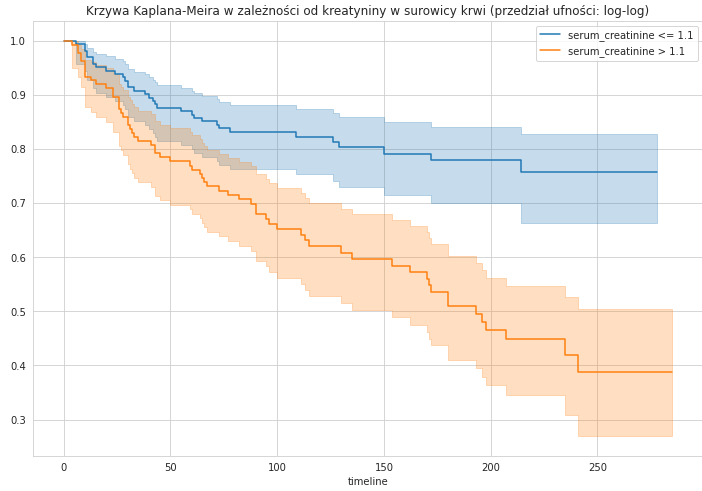


Wykonany test wskazuje na to, że należy odrzucić hipotezę zerową mówiącą o równości obu funkcji trwania. Są one różne, a zmienna wiek jest istotnym jej predyktorem. Test daje ten sam wniosek przy poziom istotności alfa=0.05 zarówno dla podziału na grupy ze względu na 60 jak i 70 lat (powyższy wynik jest dla podziału o limicie 70 lat).

Trzecią różnicującą zmienną jest poziom kreatyniny w surowicy krwi, który podczas eksploracji danych wydawał się prawdopodobnym predyktorem funkcji trwania. 

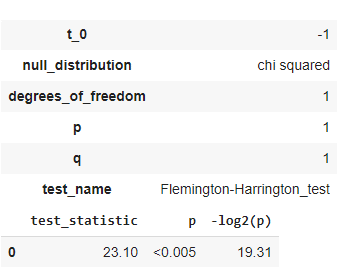


Jej rozkład zaprezentowany powyżej wskazuje na dużą zmienność, oraz bardzo długi prawy ogon rozkładu – istnieje znaczna ilość obserwacji odstających.



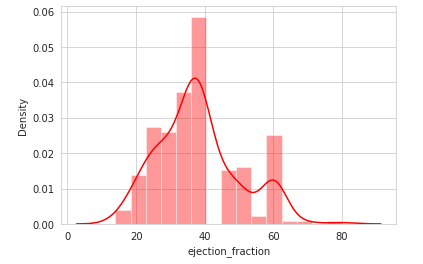
Dzieląc funkcje trwania w podziale na obserwacje poniżej oraz powyżej średniej wartości tej zmiennej, łatwo jest zauważyć znaczną różnicę w funkcji trwania na nie korzyść pacjentów o wysokim stężeniu kreatyniny w surowicy krwi.

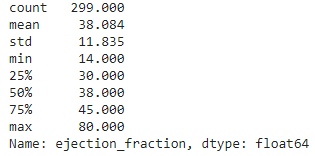
Test użyty do sprawdzenia równości funkcji trwania w tym przypadku to test Fleminga-Harrisona o wartościach p=1 i q=1, który daje większe wagi dla funkcji trwania w połowie okresu obserwacji – w tym miejscu wizualnie wydaje różnica pomiędzy grupami wydaje się zwiększać w wyższym tempie.



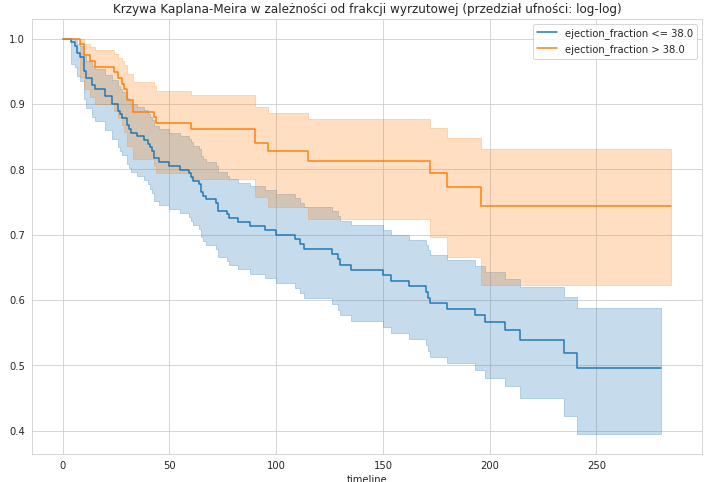
Statystyka testowa oraz p-value wskazują, że należy odrzucić hipotezę zerową o równości funkcji trwania. Poziom kreatyniny w surowicy krwi jest istotnym predyktorem funkcji przeżycia.

Ostatnią badaną zmienną objaśnianą jest frakcja wyrzutowa, która także w wyniku analizy wstępne daje oznaki bycia potencjalnym predyktorem.

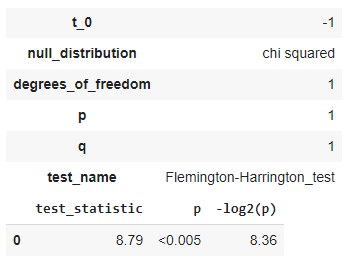




Znaczna część obserwacji oscyluje około średniej (oraz niemal równej mediany), jednak są też obserwacje w obie strony rozkładu, przy czym w prawym ogonie, maksymalne, są bardziej oddalone od średniej niż te minimalne. Odchylenie standardowe stanowi około 31% średniej arytmetycznej

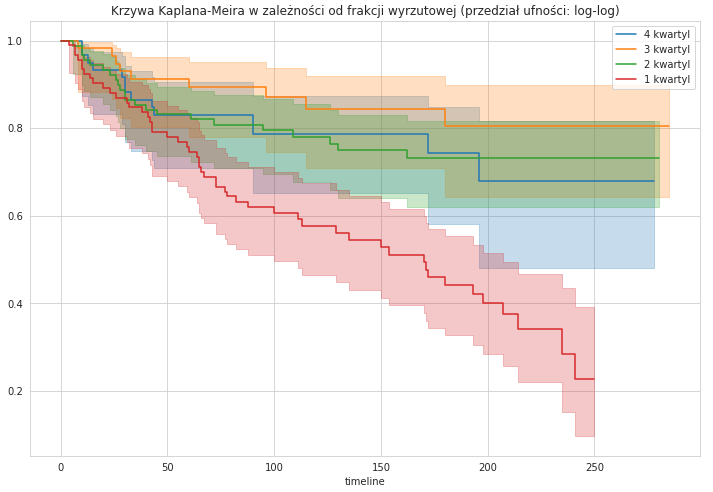


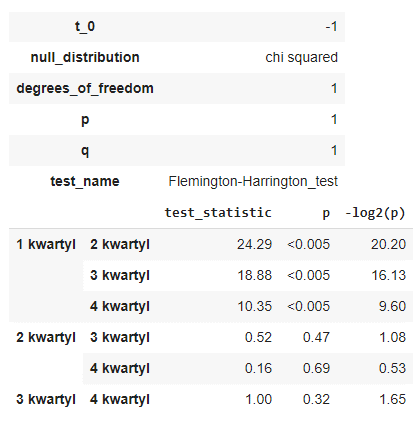
Krzywa Kaplana-Meiera wskazuje na różniące się funkcje przeżycia wśród grup podzielonych ze względu na średnią wartość zmiennej. Największa różnica w dynamice funkcji zdaje się być pośrodku zmiennej czasu obserwacji, a więc ponownie zostanie wykorzystany test Fleminga-Harringtona (1, 1).



Ponownie należy odrzucić hipotezę zerową. Frakcja wyrzutowa jest istotnym predyktorem funkcji przeżycia.

Dodatkowo wykonany został test parami ze względu na kwartyle tej zmiennej, który wskazuje na różnice w funkcji trwania szczególnie pomiędzy 1 kwartylem a pozostałymi.



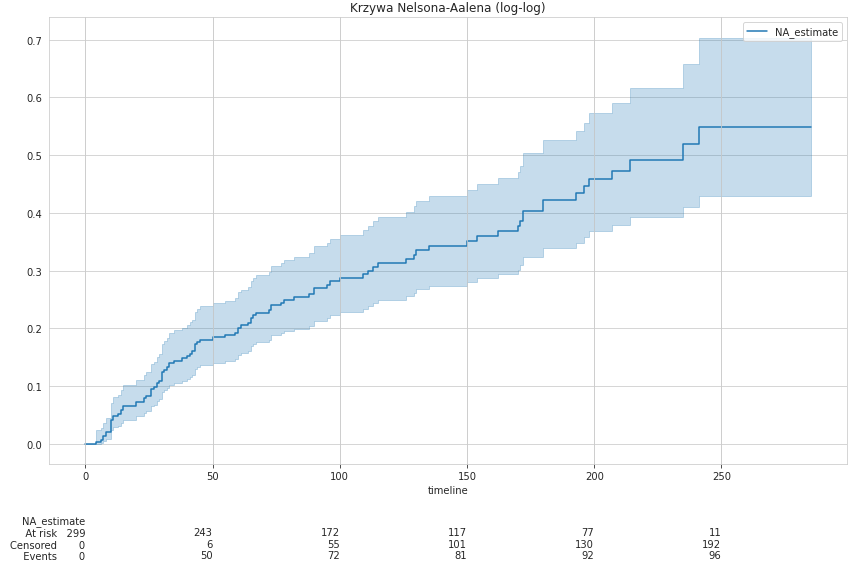


1. **Estymator Nelsona-Aalena**

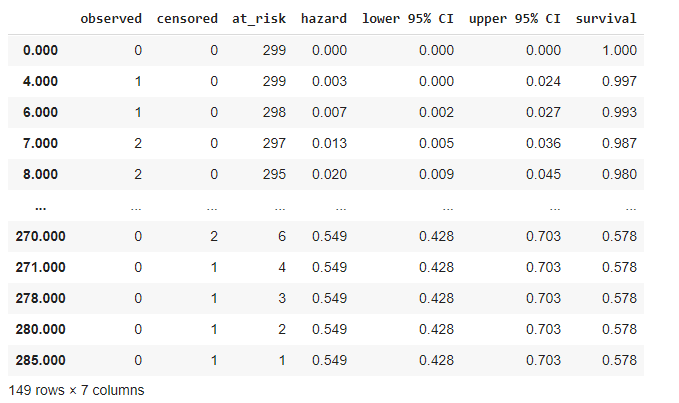
Estymator Nelsona-Aalena przyjmuje wartości opisane przez:

dla .

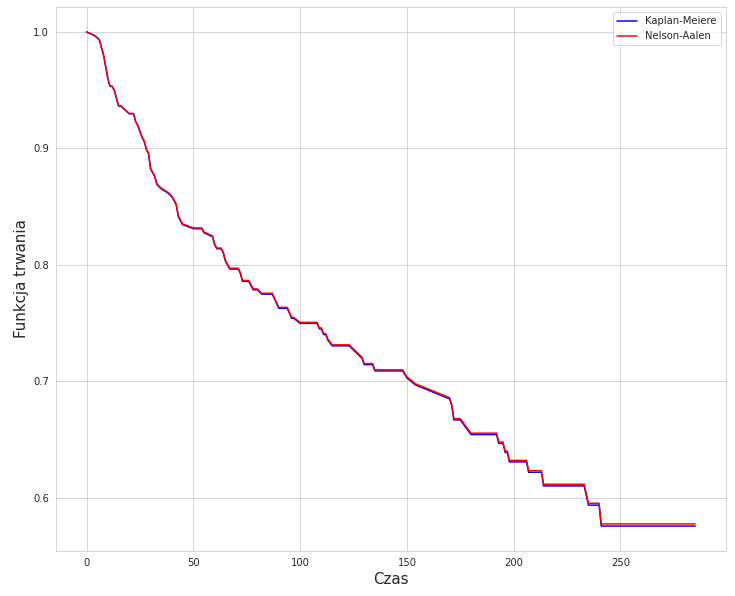
Obliczona funkcja hazardu dla analizowanego zbioru danych wygląda jak na wykresie poniżej.



Jej wartości są następujące:



Po transformacji funkcji hazardu w funkcję przeżycia oraz nanoszeniu jej na wykres razem z funkcją przeżycia obliczoną wcześniej za pomocą estymatora Kaplana-Meiera, można zauważyć, że są one w tym przypadku niemal identyczne.



Rzeczywiście, maksymalna różnica (dla ostatniego zaobserwowanego przedziału czasu) to zaledwie 0.0018.

1. **Model regresji hazardu PH Coxa**

Model regresji hazardu PH Coxa przyjmuje postać:

,

gdzie jest bazową funkcją hazardu obliczaną nieparametrycznie na podstawie estymatora funkcji przeżycia, a jest funkcją regresji o postaci:

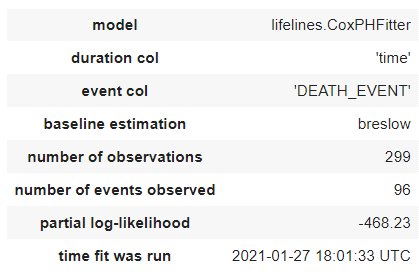
,

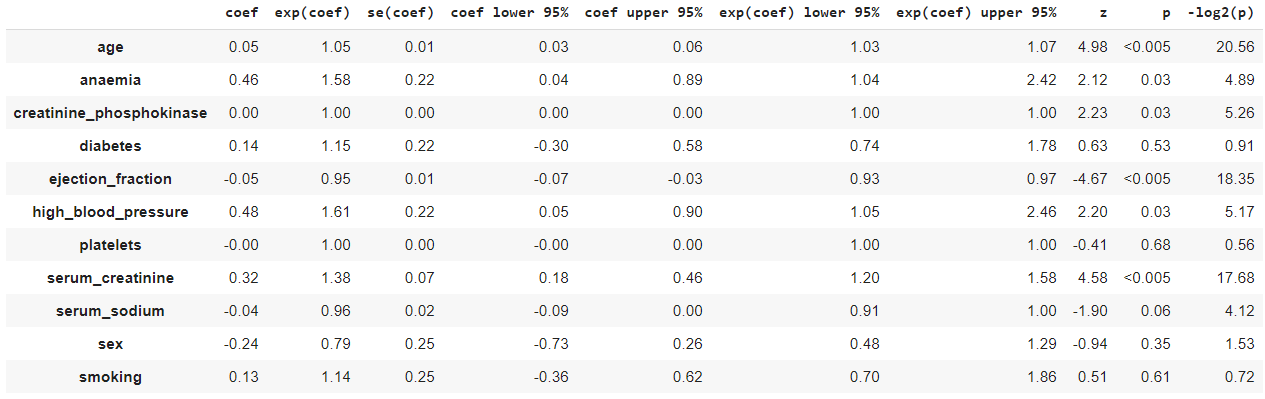
dla której parametry są parametrami strukturalnymi dla odpowiednich zmiennych objaśniających .

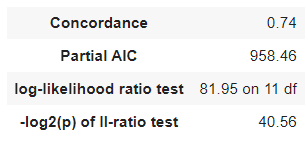
Najpierw oszacowany został model ze wszystkimi zmiennymi objaśniającymi. Parametry strukturalne modelu zostaną zinterpretowane w dalszej części badania dla ostatecznego estymowanego modelu.

W poniżej przedstawionym modelu Coxa większość zmiennych jest statystycznie nieistotna. Należy zauważyć, że analizowane wcześniej zmienne: poziom serotoniny w surowicy krwi, wiek i frakcja wyrzutowa są istotnymi predyktorami według wstępnego modelu.

Model 1:



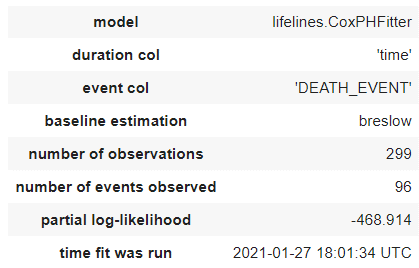


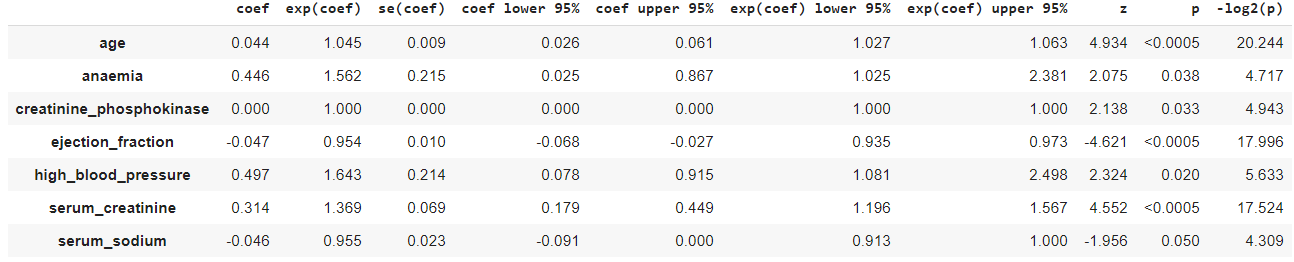


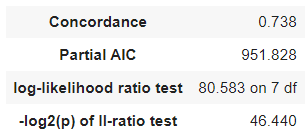
*\*bardziej czytelna wersje tabel w google colab notebook wysłanym razem z pracą.*

Po odrzuceniu zmiennych statystycznie nieistotnych (przy poziomie istotności alfa=0.05) kryterium informacyjne Akaikego przyjęło niższą wartość, co wskazuje na lepsze dopasowanie modelu. Wskaźniki -log2(p) oparty na wiarygodności także przyjął lepszą (wyższą) wartość.

Model 2:







*\*bardziej czytelna wersja tabel w google colab notebook wysłanym razem z pracą.*

Warto w tym momencie zwrócić uwagę na założenie proporcjonalnego hazardu przyjęte przez model PH Coxa, aby estymowane parametry były nieobciążone. Jest ono wyrażone jako:

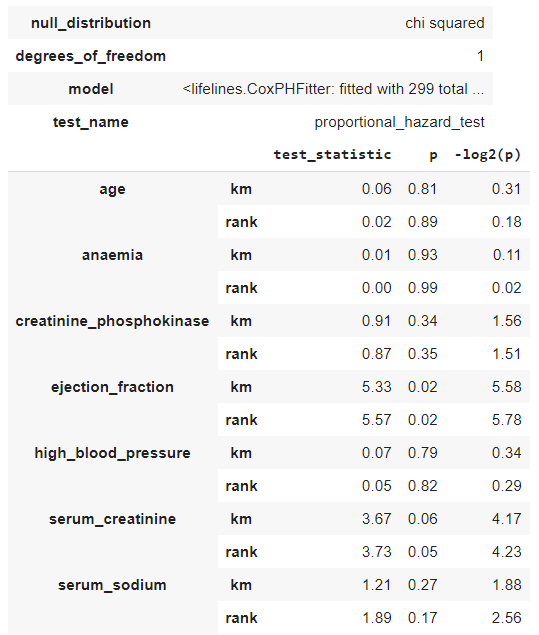
,

gdzie *a* jest wartością stałą w czasie obserwacji pacjentów. W przypadku różnych obserwacji prawdziwe musi być więc założenie:

,

gdzie współczynnik powinien być stały dla każdego czasu *t*.

Korzystając z testu Grambscha-Therneau, którego hipoteza zerowa mówi o prawdziwości proporcjonalności hazardu, sprawdzona została zmienność w czasie parametrów strukturalnych modelu 2.



Wyniki testu wskazują na to, że parametry strukturalne zmiennych frakcja wyrzutowa oraz serotonina w surowicy krwi są zmienne w czasie i łamią założenie o proporcjonalności hazardu.

Drugą metodą na potwierdzenie tych przypuszczeń jest wizualizacja wyskalowanych reszt Schoenfielda bazujących na częściowej wiarygodności. Reszty Schoenfielda wyrażone są wzorem:

,

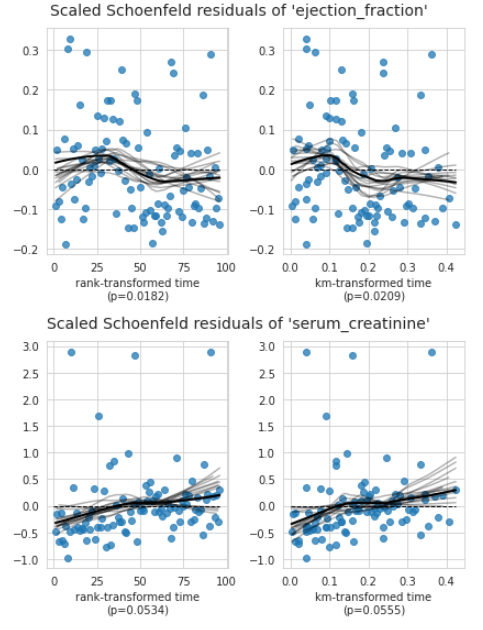
Dla jednostki *i* oraz zmiennej . Wyskalowane reszty Schoenfielda opisane są:

,

gdzie, jest liczbą zdarzeń w próbie o *n* elementach.

Reszty te są różnicą w momencie między empirycznymi wartościami zmiennych objaśniających oraz wartościami oczekiwanymi pod warunkiem, że jednostek jest narażonych. Aby spełnić założenie o proporcjonalności hazardu, wyskalowane reszty Schoenfielda powinny być bliskie 0 w każdym okresie czasu.

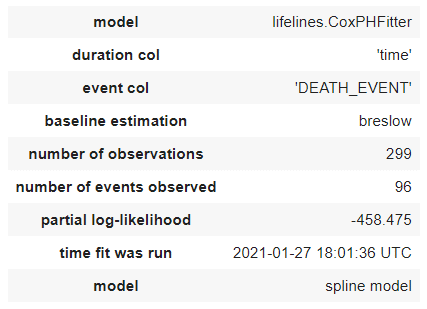
Wykresy tych reszt dla zmiennych frakcji wyrzutowej oraz serotoniny w surowicy krwi odbiegają od wartości 0:

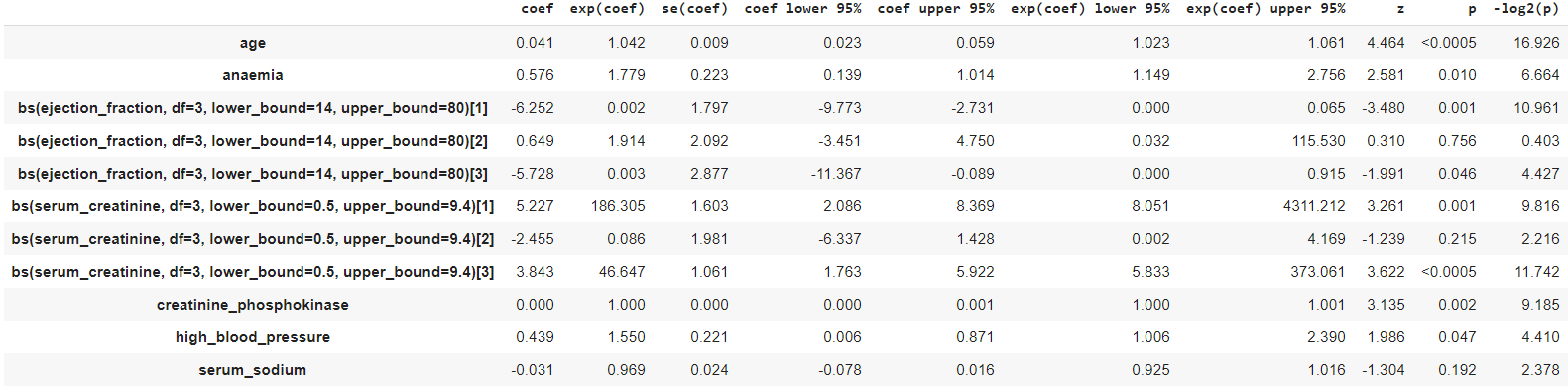


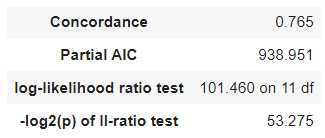
Aby skorygować te założenie do modelu zostały dodane sklejone zmienne (*B-splines*) dla frakcji wyrzutowej i serotoniny w surowicy krwi. B-spline to podział zakresu zmiennej na mniejsze przedziały, dzięki czemu możliwe jest uzyskanie nieliniowej zależności poprzez kombinację liniową tych pomniejszych funkcji bazowych.

Dzięki metodzie B-spline skalowane reszty Schoenfielda wynoszą w przybliżeniu 0 i spełnione jest założenie o proporcjonalności hazardu, jednak w zamian za to model jest znacznie mniej interpretowalny, szczególnie zważywszy, że było to bardzo istotne zmienne. Model 3 natomiast pomimo utraty interpretowalności wydaje się być lepiej dopasowany ze względu na kryterium informacyjne Akaike. Można założyć, że ma on lepszą moc predykcyjną, jednak potrzebne by były kolejne testy i miary predykcji aby to potwierdzić.

Model 3:

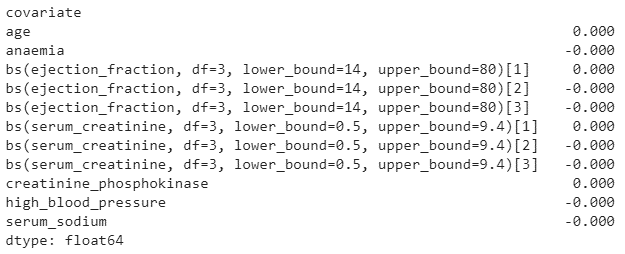






*\*bardziej czytelna wersja tabel w google colab notebook wysłanym razem z pracą.*

Wartości oczekiwane skalowanych błędów Schoenfielda modelu 3:



1. **Interpretacja najlepszego modelu**

Średni hazard bazowy dla wszystkich punktów w czasie (dla których występowały wydarzenia lub obserwacje cenzurowane) modelu 3 wynosi 0.003 – oznacza to, że zdecydowanie większą część szacowanego hazardu stanowi efekt zmienności zmiennych objaśniających.

Interpretacja parametrów strukturalnych modelu:

- Wzrost wieku pacjenta o rok wpływa na zwiększenie hazardu średni o około 4.2 %. Ceteris Paribus.

- Spadek erytocytów lub hemoglobiny u pacjenta zwiększa hazard pacjenta w danym okresie o około 78% względem pacjentów, u których nie występuje ten spadek (względem grupy referencyjnej. Ceteris Paribus.

- wzrost enzymu CPK we krwi o 1 mcg/L zwiększa hazard średnio o mniej niż 0.1%. Ceteris Paribus. Zmienna zdaje się wpływać bardzo nieznacznie na wartość funkcji hazard, pomimo to zmienna jest statystycznie istotna. Jest tak ze względu na jej duży zakres (minimum 23, maksimum powyżej 7000).

- Pacjenci z wysokim ciśnieniem krwi mają hazard średni o około 55% wyższy w tym samy okresie czasu niż pacjenci bez podwyższonego ciśnienia krwi. Ceteris Paribus.

- Zwiększenie poziomu sodu we krwi o 1 mEq/L zmniejsza wartość hazardu średnio o około 3%. Ceteris Paribus.

- Zmienna frakcja wyrzutowa jest trudna do interpretacji ze względu na transformacje przy pomocy b-spline, jednak biorąc pod uwagę na wcześniejszą analizę krzywych ze względu na tą zmienną, wiemy, że pacjenci o niskich wartościach są najbardziej narażeni, o najwyższych wartościach także są bardziej narażeni od tych w okolicy mediany i 3 kwartyla. Parametry przy zmiennych bazowych funkcji sklejanej oraz ich istotność zdają się to potwierdzać.

- Kreatynina w surowicy krwi także nie ma prostej interpretacji przez transformacje funkcją sklejaną. Podobnie, we wcześniejszej analizy krzyw przeżycia dowiedliśmy, że pacjenci z wyższą wartością tej zmiennej mają mniejszą wartość funkcji przeżycia, a więc większą wartość funkcji hazardu.

Bazując interpretacje zmiennych „frakcja wyrzutowa” oraz „kreatynina w surowicy krwi” na modelu 2 (który nie spełniał założenia proporcjonalnego hazardu i miał większy wynik kryterium AIC) brzmią one:

- Wraz ze wzrostem frakcji wyrzutowej o 1 punkt procentowy, wartość hazardu maleje średnio o około 4%. Ceteris Paribus.

- Wraz ze wzrostem kreatyniny w surowicy krwi o 1 mg/dL, wartość hazardu rośnie średnio o około 37%. Ceteris Paribus.