Politechnika Wrocławska

Wydział Matematyki
Kierunek studiów: Matematyka i Statystyka
Specjalność: Statystyka i analiza danych

Praca dyplomowa – licencjacka

METODA TIME-DEPENDENT-ROC W ANALIZIE PRZEŻYCIA

Aleksandra Grzeszczuk

słowa kluczowe:
Analiza przeżycia, funkcja przeżycia, funkcja hazardu, krzywa ROC, krzywa ROC zależna od czasu, rozkład Birnbauma–Saundersa, rozkład log-normalny

krótkie streszczenie:
W pracy zostaną pokazane metody estymacji krzywej ROC w przypadkach, gdy funkcje czułości i swoistości zależą od czasu. Sformułowane zostaną podobieństwa oraz różnice w metodach estymacji krzywej ROC w przypadku zależnym od czasu oraz bez tej zależności. Przedstawimy warunki stosowalności podanych metod. Przeprowadzone zostaną symulacje opisanych metod i porównanie uzyskanych wyników numerycznych i analitycznych. Implementacja algorytmów zostanie wykonana w środowisku programistycznym R.

Opiekun pracy dyplomowej	dr Marek Skarupski	.....	.....
	Tytuł/stopień naukowy/imię i nazwisko	ocena	podpis

*Do celów archiwalnych pracę dyplomową zakwalifikowano do:**

a) kategorii A (akta wieczyste)

b) kategorii BE 50 (po 50 latach podlegające ekspertyzie)

** niepotrzebne skreślić*

pieczątka wydziałowa

Wrocław, rok 2022

Wrocław University
of Science and Technology

Faculty of Pure and Applied Mathematics
Field of study: Mathematics and Statistics
Specialty: Statistics and Data Analysis

Bachelor's Thesis

TIME-DEPENDENT-ROC METHOD IN SURVIVAL ANALYSIS

Aleksandra Grzeszczuk

keywords:
Survival analysis, survival function, hazard funcion, ROC Curve, Time-dependent ROC Curve, Birnbaum–Saunders distribution, log-normal distribution

short summary:
This thesis presents methods of ROC curve estimation in cases where the sensitivity and specificity functions depend on time. The similarities and differences in the ROC curve estimation methods with and without time-dependent dependence will be formulated. The conditions of applicability of the given methods will be presented. The described methods will be simulated and the obtained numerical and analytical results will be compared. The implementation of the algorithms will be made in the R programming environment.

Supervisor	dr Marek Skarupski	.....	.....
	Title/degree/name and surname	grade	signature

*For the purposes of archival thesis qualified to:**

a) category A (perpetual files)

b) category BE 50 (subject to expertise after 50 years)

** delete as appropriate*

stamp of the faculty

Wrocław, 2022

Spis treści

Rozdział 1

WSTĘP

1.1 Wprowadzenie do tematu

Epidemia COVID - 19 wywołana przez koronawirusa SARS - CoV - 2 rozpoczęła się 17 listopada 2019 roku w mieście Wuhan w środkowych Chinach, zaś 11 marca 2020 roku została uznana przez Światową Organizację Zdrowia (WHO) za pandemię[1]. Przez cały ten okres powstało wiele artykułów zawierających analizę przeżycia pacjentów chorujących na COVID - 19.

Salinas - Escudero i in. [38] zastosowali analizę przeżycia, aby zbadać wpływ COVID - 19 na populację meksykańską. Objęła ona wszystkie potwierdzone przypadki choroby zawarte w zbiorze danych opublikowanych przez System Nadzoru Epidemiologicznego Wirusowych Chorób Oddechowych Ministerstwa Zdrowia Meksyku. Na podstawie tej analizy wykreślili krzywe Kaplana - Meiera oraz skonstruowali model proporcjonalnego hazardu Coxa.

Lu i in. [29] zajęli się analizą, której celem była ocena klinicznych cech i wyników u pacjentów z koronawirusem w 2019 roku. Chcieli oni także pomóc klinicystom w przeprowadzeniu prawidłowego leczenia i ocenie rokowań pacjentów.

Purkayastha i in. [36] opracowali system uczenia maszynowego opartego na radiomice[2] w celu przewidywania ciężkości choroby COVID - 19 i przyszłego pogorszenia stanu krytycznej choroby za pomocą różnych zmiennych klinicznych. Stworzyli oni model do przewidywania ciężkości oraz model czasu do zdarzenia, do przewidywania progresji do choroby krytycznej. W celu określenia wydajności modelu obliczali obszar pod krzywymi ROC - AUC, obliczali wskaźnik krytyczny (c) zależny od czasu, a następnie dokonywali wizualnej interpretacji.

Hasab [19] zbadał wpływ testów RT - PCR na kontrolę pandemii COVID - 19 i ważność RT - PCR jako predyktora choroby. Dane zostały zebrane przy użyciu danych wtórnych. W badaniach uwzględniono wszystkie przypadki i zgony w raportach sytuacyjnych WHO. Przeprowadzono analizę przeżycia w celu określenia skumulowanego proporcjonalnego przeżycia COVID - 19 w Egipcie. Przeprowadzono również analizę krzywej ROC, która

[1] https://pulsmedycyny.pl/who-oglosilo-pandemie-covid-19-co-to-oznacza-984790

[2] Radiomika metodą, która wyodrębnia dużą liczbę cech z obrazów medycznych przy użyciu algorytmów charakterystyki danych.

została wykorzystana do zbadania wydajności diagnostycznej RT - PCR.

Tematem niniejszej pracy są zagadnienia związane z analizą przeżycia. Stanowi ona zbiór procedur statystycznych do analizy danych, w których interesującą nas zmienną wynikową jest czas do wystąpienia pewnego określonego zdarzenia. Z matematycznego punktu widzenia, analiza przeżycia opiera się na rachunku prawdopodobieństwa, statystyce, różnych metodach optymalizacji oraz na procesach stochastycznych. Obecnie jej techniki znajdują szerokie zastosowanie głównie w badaniach klinicznych, ale także w ekonomicznych czy demograficznych.

Celem pracy jest sformułowanie podobieństw oraz różnic w metodach estymacji krzywej ROC w przypadku zależnym od czasu, oraz bez tej zależności.

Szeroko omówione zostały parametryczne i nieparametryczne metody estymacji funkcji przeżycia, testy statystyczne w analizie przeżycia oraz nieparametryczny model Coxa. Przykłady zastosowań oraz dokładne omówienie wyników znajduje się w rozdziale *Analiza przeżycia na podstawie danych z biblioteki survival w R.*

Praca została podzielona na cztery części. Pierwsza z nich opisuje teorię modeli przeżycia i krzywych ROC, dane cenzurowane oraz krzywe ROC zależne od czasu, a także rozkłady probabilistyczne czasów życia. Każde zagadnienie opiera się na szczegółowo dobranej literaturze. Również w tym rozdziale zostały opisane podstawowe charakterystyki analizy przeżycia - począwszy od definicji, po przykłady, funkcje przeżycia aż do funkcji hazardu.
W rozdziale drugim omówiony został najpopularniejszy estymator analizy przeżycia - Kaplana Meiera, wspomniane wcześniej metody estymacji funkcji przeżycia, testy statystyczne i nieparametryczny model Coxa.
W rozdziale trzecim ukazano podstawowe fakty dotyczące rozkładów log - normalnych oraz Birnbauma - Saundersa. Rozdział czwarty zawiera analizę danych zobrazowanych na podstawie modeli pochodzących z rozkładów logarytmicznie normalnych oraz Birnbauma - Saundersa.

1.2 Przegląd literatury

1.2.1 Modele przeżycia i krzywe ROC

Heagerty oraz Zheng [21] zaproponowali nowe podsumowania dokładności zależne od czasu w oparciu o specyficzne dla czasu wersje czułości i swoistości obliczonych dla zbiorów ryzyka. Połączone zostały podsumowania dokładności z wcześniejszą miarą zgodności będącą wariantem τ Kendalla (*Kendall's tau)*[3]. Pokazane zostało nowe wykorzystanie standardowych danych wyjściowych regresji Coxa do uzyskania oszacowań czułości i specyficzności zależnej od czasu oraz krzywych ROC zależnych od czasu.

Wcześniejsze badania koncentrowały się na rozszerzeniu proporcji zmienności wyjaśnianej przez współzmienne, lub R^2, do modeli ocenzurowanych. Schemper i Henderson [40] określili nową miarę proporcji zmienności możliwych ocenzurowanych czasów przeżycia wyjaśnioną przez dany model proporcjonalnych hazardów. Zasugerowali miarę dokładności predykcyjnej i przyrostów dokładności predykcyjnej, które można uznać za narzędzia do ilościowego określania wiedzy, na przykład medycznej na temat przebiegu choroby. Celem tych miar jest dostarczanie dodatkowych informacji podsumowujących, uzupełniających zwykłą kontrolę dopasowanych krzywych przeżycia dla różnych grup prognostycznych.

Heagerty i Zheng [21] wykazali związek między metodami krzywych ROC zależnych od czasu a klasycznymi podsumowaniami zgodności, takimi jak τ Kendalla lub "*c Index*". Harell oraz Lee [17] zdefiniowali "*c Index*" jako odsetek wszystkich użytecznych par pacjentów, w których prognozy oraz wyniki są zgodne. Mierzy on informacje predykcyjne pochodzące z zestawu zmiennych predykcyjnych w modelu. Oszacowując czas do śmierci, *c* oblicza się, biorąc pod uwagę wszystkie możliwe pary pacjentów, w których przynajmniej jeden zmarł. Jeśli przewidywany czas przeżycia jest większy dla pacjenta żyjącego dłużej, to stwierdzamy, że przewidywania dla tej pary są zgodne z wynikami. Jeżeli jeden pacjent zmarł, a wiadome jest, że drugi przeżył co najmniej czas przeżycia pierwszego, zakładami, że drugi pacjent przeżyje pierwszego. "*c Index*" szacuje prawdopodobieństwo zgodności między zmienną przewidywaną a obserwowaną, gdzie wartość 0.5 oznacza brak predykcyjnego rozróżniania, zaś wartość 1 określa doskonałe oddzielenie pacjentów z różnymi wynikami.

Pantoja - Galicia [35] w kontekście choroby Alzheimera zilustrował powiązania między definicjami czułości i swoistości zależnej od czasu z miarami zgodności. Ustanowił także nowe powiązania za pomocą nowych miar zgodności globalnej, a także zbadał związek między takimi miarami a odpowiadającymi im AUC zależnymi od czasu. Gneiting i Waltz [14] zaproponowali wykorzystanie klasycznej analizy ROC w danych z biomedycyny i meteorologi, gdzie miary oparte na rangach dostarczają nowych informacji w porównaniu z prognozą pogody w WeatherBench[4], dotyczących wydajności predykcyjnej splotowych sieci neuronowych i fizyczno - numerycznych modeli do przewidywania pogody.

[3]W statystyce współczynnik τ Kendalla jest statystyka używaną do pomiaru powiązania porządkowego (*ordinal assciation* pomiędzy dwiema mierzonymi wielkościami. Test τ jest nieparametrycznym testem hipotezy na zależność statystyczną na podstawie współczynnika τ.

[4]WeatherBench jest zestawieniem danych porównawczych do prognozowania pogody na podstawie danych

Diaz - Coto, Martinez - Camblor oraz Perez - Fernandez [12] skoncentrowali się na implementacji krzywych ROC w programie `R` (pakiet `smoothROCtime`) w celu płynnego oszacowania zależnych od czasu krzywych ROC. Zaprezentowany został teoretyczne powiązanie krzywych C/D oraz I/D ROC zależnych od czasu, które poprzez łączny rozkład markera i zmiennych czasu do zdarzenia, podało metodę aproksymacji wrażliwości kumulacyjnej/incydentalnej i swoistości dynamicznej.

Bensal i Heagherty [2] opracowali przegląd nowoczesnych metod statystycznych do oceny zmieniającej się w czasie dokładności podstawowego markera prognostycznego. Porównali podejścia uwzględniające zdarzenia skumulowane oraz incydenty, a tajże powszechne podejście wykorzystujące współczynniki ryzyka uzyskane z regresji proporcjonalnych hazardów Coxa z nowszymi podejściami wykorzystującymi krzywe ROC zależne od czasu. Odkryli, że zmienne w czasie HR przy użyciu lokalnej liniowej estymacji, wyraźnie ujawniły trendy czasowe poprzez bezpośrednie oszacowanie związku w każdym punkcie czasowym t, w porównaniu z analizami charakterystycznymi uśrednionymi w czasie $\geq t$.

Li, Ning oraz Feng [27] zainteresowani byli opracowaniem, oraz walidacją narzędzia do przewidywania ryzyka w celu identyfikacji przyszłych przypadków raka płuc poprzez zintegrowanie informacji demograficznych, charakterystyki choroby i danych związanych z paleniem. Biorąc pod uwagę długi okres utajnienia nowotworu, chcielibyśmy, aby narzędzie prognostyczne osiągnęło skuteczność rozróżniającą, która nie słabnie z czasem. Zaproponowali procedury szacowania i wnioskowania w celu kompleksowej oceny zarówno ogólnej dyskryminacji predykcyjnej jak i wzorca czasowego szacowanej reguły przewidywania. Zaproponowane metody można stosować z powszechnie stosowanymi modelami cenzurowanych regresji, na przykład modelu proporcjonalnego hazardu Coxa, czy przyśpieszonego czasu awarii. Omówione metody oferują narzędzia informacyjne służące między innymi do porównywania wyników dyskryminacji między różnymi modelami.

Choroba Alzheimera jest nieuleczalną i postępującą chorobą, zaczynając się od łagodnego upośledzenia funkcji poznawczych, które z czasem się pogarszają. Zbadanie wpływu pogorszenia funkcji poznawczych pacjentów na czas do rozwinięcia się w chorobę Alzheimera i uzyskanie wiarygodnego modelu diagnostycznego jest kluczowe dla oceny rokowania oraz wczesnego leczenia. Kang, Pan, Song [24] opracowali nowy model mający zaradzić różnym niedociągnięciom takim jak niepełny pomiar funkcji poznawczych czy też przeoczenie subtelnej trajektorii upośledzenia poznawczego pacjentów. Przyjęto model dynamicznej analizy czynnikowej, aby w sposób kompleksowy scharakteryzować upośledzenie funkcji poznawczych za pomocą wielu miar poznawczych, a także zaproponowano model współczynników losowych opartych na splajnie w celu ujawnienia prawdopodobnie nieliniowej trajektorii pogorszenia funkcji poznawczych pacjentów. Rozważony został tutaj także model proporcjonalnego hazardu w celu zbadania wpływu niezmiennych w czasie markerów i zmiennego w czasie upośledzenia funkcji poznawczych zagrożenia chorobą Alzheimera.

1.2.2 Dane cenzurowane i krzywe ROC zmienne w czasie

Krzywe ROC są popularną metodą prezentowania wrażliwości i swoistości ciągłego markera diagnostycznego X dla binarnej zmiennej choroby D (szersze omówienie krzywych ROC znajduje się w rozdziale 2.5 *Krzywe ROC*). Wiele skutków choroby jest czasowo za-

leżne ($D(t)$), więc krzywe ROC zależne od czasu są zazwyczaj bardziej odpowiednią metodą do badania tych zależności. Heagerty, Lumley oraz Pepe [20] zaproponowali podsumowanie potencjału markera X mierzonego na początku (czyli w czasie $t = 0$), poprzez obliczenie krzywych ROC dla skumulowanej liczby zachorowań (bądź zgonów) w czasie t. Oznaczamy to jako ROC(t) (*Time - Dependent ROC*). Typową komplikacją dotyczącą analizy danych przeżycia, jest możliwość występowania danych cenzurowanych, zaproponowane zostały zatem dwa estymatory krzywej ROC, które mogą uwzględniać takowe dane. Prosty estymator Kaplana - Meiera nie zawsze gwarantuje konieczny warunek, dotyczący monotoniczności w X funkcji czułości i swoistości. Alternatywny estymator, gwarantujący tę monotoniczność, oparty jest na estymatorze najbliższego sąsiada dla dwuwymiarowej funkcji rozkładu (X, T), gdzie T oznacza czas przeżycia. Akritas [1] rozważył ten problem szacowania rozkładu dwuwymiarowego wektora losowego (X, T), gdzie zmienna T może podlegać losowemu cenzurowanie. Zmienna cenzurująca C może zależeć od X, jednakże zakłada się, że T oraz C są warunkowo niezależne. Oszacowanie rozkładu dwuwymiarowego uzyskuje się poprzez uśrednienie oszacowań rozkładu warunkowego T przy danym $X = x$, w zakresie wartości parametru x. Korzystając z zaproponowanego estymatora, uzyskuje się rozszerzenie estymatora metodą najmniejszych kwadratów na ocenzurowaną regresję wielomianową danych, po czym ustala się jego asymptotyczną normalność. Heagerty, Lumley oraz Pepe [20] korzystając z tych zależności, przedstawiają przykład, w którym krzywa ROC(t) służy do porównania standardowego oraz zmodyfikowanego pomiaru cytometrii przepływowej do przewidywania przeżycia po wykryciu raka piersi, a także analizują wpływ modyfikacji kryteriów kwalifikowalności dla wielkości próbki i mocy w badaniach profilaktyki HIV.

Rak piersi jest najczęstszym nowotworem złośliwym o wysokiej heterogeniczności u kobiet, dla którego prognozy nadal pozostają słabe. Lu i in. [30] z publicznych baz danych zebrali listę odpowiednich danych kliniczno - patologicznych i podzielili pacjentów na kohortę treningową oraz kohortę walidacyjną. Wykorzystali jednoczynnikową analizę regresji Coxa do identyfikacji prognostycznych genów związanych z ferroptozą [5], zaś późniejsza analiza wielowymiarowa przeanalizowała ważne geny w celu ustalenia modelu prognostycznego. Do walidacji modelu w kohortach wewnętrznych i zewnętrznych wykorzystano krzywe ROC.

Noda [33] skupił się na ocenie użyteczności analizy histogramu (HA) wartości pozornego współczynnika dyfuzji do przewidywania całkowitego przeżycia (OS) u pacjentów z gruczolakorakiem przewodowym trzustki i skorelowania z patologicznie ocenioną masywną martwicą wewnątrz guzową. Do retrospektywnego badania włączono 39 pacjentów, którzy zostali poddani rezonansowi magnetycznemu. Stosując analizę regresji proporcjonalnej Coxa skorygowaną o wiek, analizę krzywej ROC zależnej od czasu oraz oszacowanie Kaplana - Meiera autorzy ocenili związek między parametrami HA a OS.

1.2.3 Rozkłady probabilistyczne czasów życia

Rozkłady probabilistyczne o nośnikach na dodatniej półprostej rzeczywistej są naturalnym wyborem przy modelowaniu czasu życia. Wśród nich szczególną rolę odgrywa rozkład wykładniczy ze względu na własność braku pamięci. W kontekście analizy przeżycia szczególną uwagę skupimy na rozkładach pochodzących od rozkładu normalnego, jak np. seminormalny czy log-normalny. Dołączymy jeszcze rozkład Birnbauma - Saundersa.

[5] Ferroptoza jest formą śmierci komórkowej, która charakteryzuje się rosnącym poziomem nadtlenków lipidów. Została ona zdefiniowana jako mechanizm śmierci komórek w na przykład sepsie.

Birnbaum i Saunders [6] przedstawili nową dwuparametrową rodzinę rozkładów długości życia wyprowadzoną z modelu zmęczenia. Nazywana jest ona rozkładem trwałości zmęczeniowej. Wyprowadzenie to wynika z rozważań teorii odnowy dla liczby cykli potrzebnych do wymuszenia rozszerzenia pęknięcia zmęczeniowego wartości krytycznej. Problem estymacji dla nowej dwuparametrowej rodziny rozkładów długości życia został kontynuowany w [5]. Uzyskali oni oszacowanie prawdopodobieństwa obu parametrów i podali oraz zbadali iteracyjne procedury obliczeniowe. Przedstawili proste oszacowania mediany czasu życia, a następnie porównano je z oszacowaniem największego prawdopodobieństwo. Wykazano także, że asymptotyczny rozkład tego oszacowania mieści się w tej samej klasie rozkładów, co same obserwacje.

Raaijmakers [37] wyprowadził wzór na funkcję gęstości prawdopodobieństwa czasu życia dla układu identycznych jednostek. Czas życia systemy jest tutaj sumą czasów życia niezależnych jednostek, gdy ich czasy życia są rozłożone niezależnie zgodnie z zasadami zaproponowanymi przez Birnbauma - Saundersa.

Saunders [39] wykazał, że niektóre ze znanych właściwości określonej rodziny rozkładów, wyprowadzone jako model zniszczenia zmęczeniowego są prawidłowe dla klasy rodzin. Wynika to z faktu, że każda zmienna w odpowiednio wyskalowanej klasie ma taki sam rozkład, jak jej odwrotność. Własność ta wpływa między innymi na oszacowanie parametru skali i umożliwia rozwiązanie dwóch ważnych problemów praktycznych.

Bhattacharyya oraz Fries [4] opisali odwrotny rozkład Gaussa, który ma wyraźną przewagę nad rozkładem Birnbauma - Saundersa pod względem dostępności procedur dla rzetelnej analizy statystycznej i wykonalności rozkładów próbkowania. Założenie wspólnego procesu zniszczenia pochodzącego z odwrotnego rozkładu Gaussa uzyskiwane jest poprzez dokładne wyprowadzenia, podczas gdy rozkład trwałości zmęczeniowej Birnbauma - Saundersa obejmuje pewne przybliżenia, przez co nie jest aż tak dokładny.

Padgett [34] rozważał szacowanie niezawodności rozkładu trwałości zmęczeniowej, gdzie parametrem skali jest mediana czasu życia. Zakładając, że parametr skali jet znany otrzymujemy estymatory Bayesa funkcji rzetelności dla rodziny odpowiednio sprzężonych oraz dla niejasnego poprzedzającego parametru kształtu Jeffreysa. Zaś gdy oba parametry są nieznane, proponuje się zmodyfikowany estymator rzetelności Bayesa, wykorzystujący estymator momentu parametru skali.

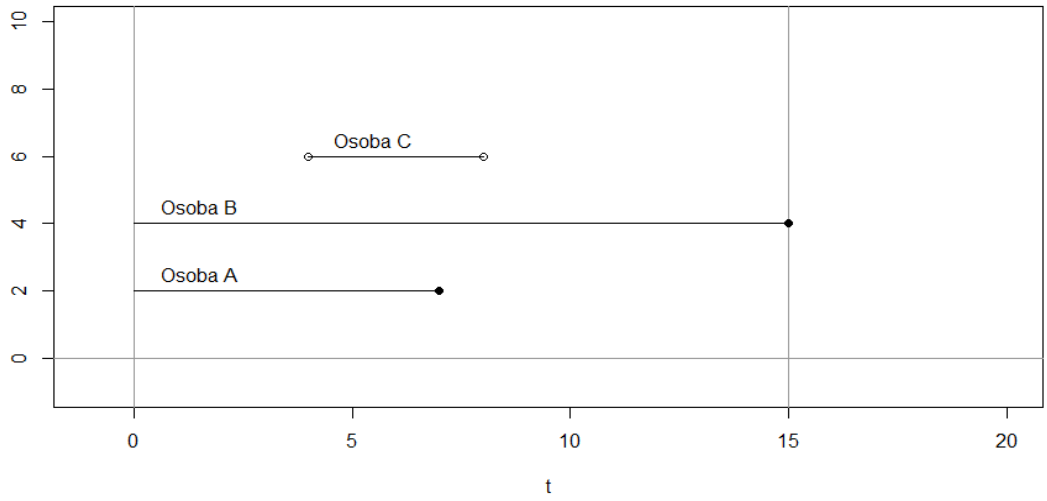
Lin oraz Yang [28] zaproponowali nową teorię statystyczną do analizy pęknięć zmęczeniowych, która jest oparta na koncepcjach mechaniki pękania procesów losowych. Skupili się oni na przypuszczalnie bardziej użytecznych informacjach o losowych czasie, w którym rozmiar pęknięcia wzrasta do określonej wartości. Biorąc pod uwagę początkowy rozmiar pęknięcia, otrzymujemy zależność rekurencyjną dla momentów statystycznych tego losowego czasu dla dość ogólnej klasy zachowań materiałów. Autorzy zilustrowali również procedurę szacowania parametrów w modelu potęgowym, wykorzystującą dane eksperymentalne niektórych próbek otworów na łączniki aluminiowe poddanych wzbudzeniu o określonym spektrum obciążenia.

1.3 Podstawowe charakterystyki analizy przeżycia

Analiza przeżycia (*survival analysis*) jest zbiorem metod statystycznych badających procesy, w których interesuje nas czas, jaki upłynie do pierwszego wystąpienia określonego zdarzenia. Tym zdarzeniem może być śmierć pacjenta, nawrót choroby, czy czas, jaki minie do ponownego aresztowania osób przebywających na zwolnieniu warunkowym (nazywamy to badaniem recydywy - powrót do przestępstwa). W analizie przeżycia zwykle odnosimy się do zmiennej czas jako czas przeżycia, ponieważ daje nam to czas, który dana osoba przetrwała. Zazwyczaj określamy to wydarzenie jako niepowodzenie, ponieważ interesującym nas wydarzeniem jest śmierć pacjenta bądź inne negatywne doświadczenie. Jednak jako czas przeżycia możemy też określić czas powrotu do pracy po planowanym zabiegu chirurgicznym. W tym wypadku niepowodzenie jest pozytywnym wydarzeniem. Metody analizy przeżycia mają szerokie zastosowanie w różnych naukach i dziedzinach życia, między innymi w ekonomii, medycynie, biologii czy nawet socjologii.

Jednym z głównych problemów w analizie przeżycia jest cenzurowanie danych (*censoring data*). Cenzurowane dane to takie, dla których zdarzenie miało miejsce przed lub po czasie obserwacji (ale nie wiadomo dokładnie kiedy). Nazywane są one odpowiednio danymi cenzurowanymi lewostronnie (jeśli zdarzenie miało miejsce przed rozpoczęciem obserwacji ($T < C$)) i prawostronnie (jeśli zdarzenie miało miejsce po czasie obserwacji ($T > C$)). Wyróżniamy także cenzurowanie przedziałowe - obserwujemy dwa punkty czasowe, między którymi wystąpiło zdarzenie.

PRZYKŁAD
Planujemy przeprowadzić 15 tygodniowy eksperyment badający, ile czasu przeżywają pacjenci po przeszczepieniu serca. Osoba A, obserwowana od początku badania umiera 7 tygodni po przeszczepie - czas jej przeżycia wynosi 7 tygodni i nie jest ocenzurowany. Osoba B obserwowana jest również od początku badania aż do jego 15 tygodniowego końca, bez uzyskania zdarzenia. Czas jej przeżycia nie jest ocenzurowany, ponieważ przeżyła cały okres naszego badania. Osoba C przystępuje do badania w 4 tygodniu i wycofuje się z niego w 8 tygodniu - czas przeżycia tej osoby jest cenzurowany po 4 tygodniach.



Rysunek 1.1: Różne rodzaje cenzurowania danych - opracowanie własne

Wprowadzimy podstawową terminologię matematyczną i notację do analizy przeżycia. Niech T będzie dowolną zmienną losową opisaną na przestrzeni probabilistycznej (Ω, F, P), określającą czas przeżycia osoby, przyjmującą wszystkie możliwe wartości nieujemne oraz t - dowolna konkretna wartość odpowiadająca zmiennej losowej T. Przykładowo, jeśli interesuje nas przeżycie danej osoby dłużej niż 5 lat po przejściu chemioterapii, to wtedy $t = 5$, zaś nasze pytanie brzmi: czy $P(T > 5)$?
Następnie $d \in \{0, 1\}$ - zmienna losowa wskazująca na porażkę ($d = 1$) bądź ocenzurowanie danych ($d = 0$).

Wprowadźmy jeszcze warunki obowiązkowo brane pod uwagę w każdej analizie przeżycia. Jest to $S(t)$ - funkcja przeżycia (*survival function*) opisana jako:

$$S(t) = P(T > t) = 1 - F(t^-) \tag{1.1}$$

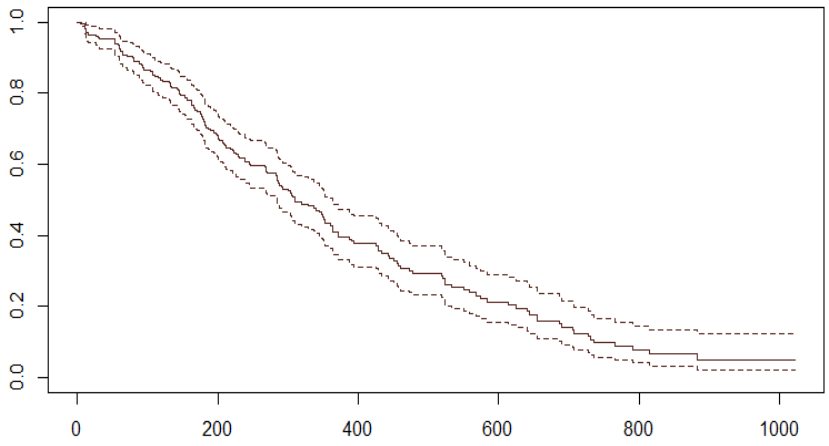
funkcja hazardu (*hazard function*) $h(t)$ zadana wzorem:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{1.2}$$

oraz skumulowany hazard (*cumulative hazard function*):

$$\Lambda(t) = \int_0^t h(u)\, du \tag{1.3}$$

Funkcja przeżycia daje nam prawdopodobieństwo, że dana osoba przeżyje dłużej niż t. Poniżej przedstawiamy wykres funkcji przeżycia dla danych `lung`[6] z biblioteki `survival`. Dane te dotyczą przeżycia pacjentów z zaawansowanym rakiem płuc.



Rysunek 1.2: Wykres funkcji przeżycia z 95% przedziałem ufności - opracowanie własne

[6]Dane pochodzące z North Central Cancer Treatment Group (https://stat.ethz.ch/R-manual/R-devel/library/survival/html/lung.html)

Przyglądając się powyższemu przykładowemu wykresowi, możemy łatwo określić następujące cechy funkcji przetrwania:

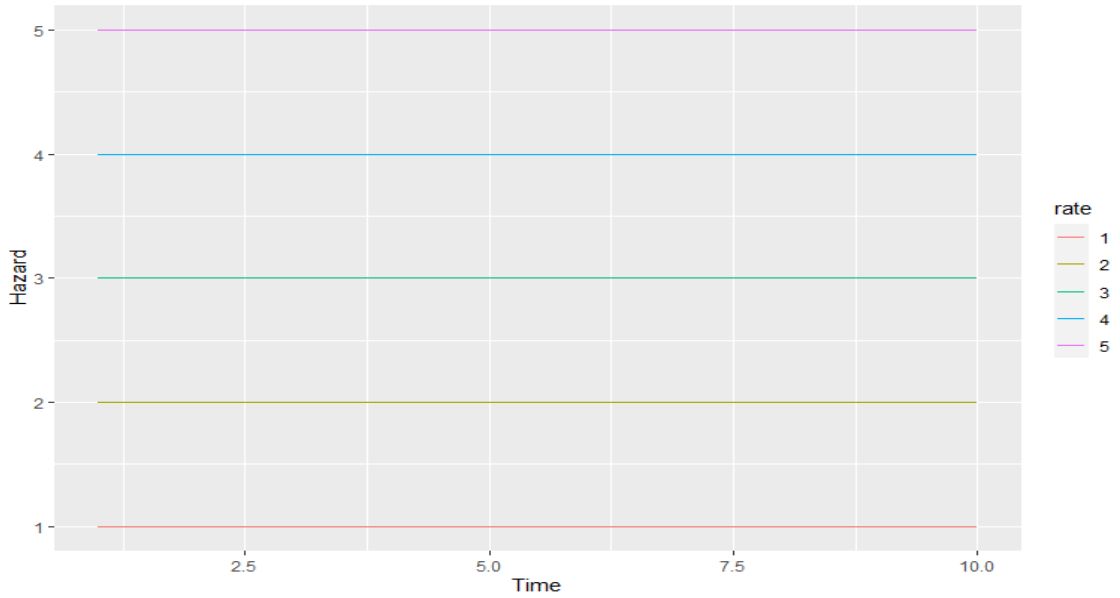
- Są to funkcje nierosnące
- Dla $t = 0$, $S(t) = S(0) = 1$, ponieważ na samym początku badania nie zaszło interesujące nas zdarzenie - zatem prawdopodobieństwa przetrwania wynosi 1.
- Dla $t = \infty$, $S(t) = S(\infty) = 0$. Zakładając, że czas obserwacji wydłużyłby się bez ograniczeń, w końcu nikt by nie przeżył - ostatecznie krzywa przeżycia musi spaść do zera.

Funkcja hazardu podaje nam chwilowy potencjał w czasie do wystąpienia zdarzenia, biorąc pod uwagę, że dana osoba przetrwała przez pewien okres czasu t. Posiada ona dwie istotne cechy:

- Jest zawsze nieujemna.
- Nie ma górnej granicy.

Kształt funkcji hazardu dostarcza nam wiele ważnych informacji jakościowych dotyczących przetrwania, dlatego na podstawie monotoniczności $h(t)$ wyróżniamy kilka klasyfikacji.

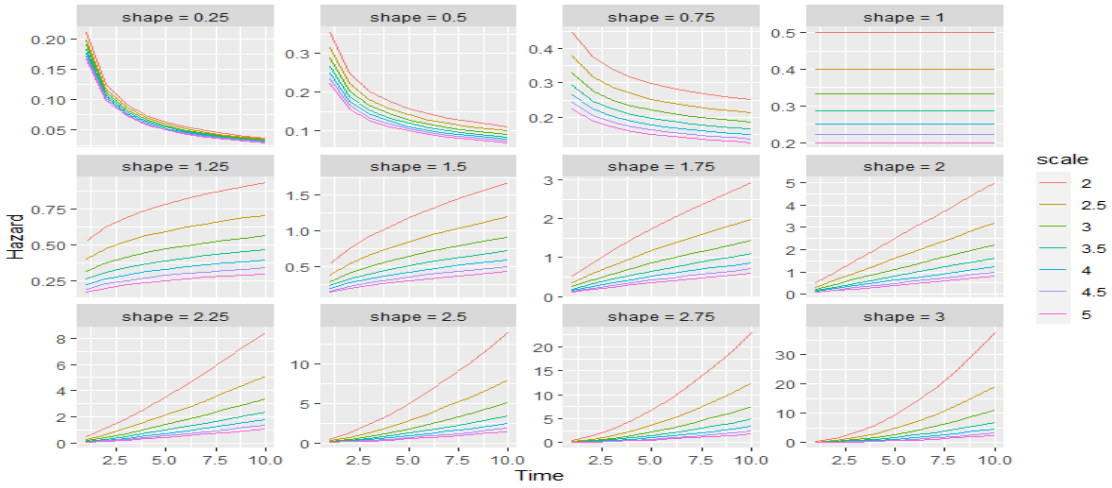
1. *Constant hazard rate* - stała funkcja hazardu



Rysunek 1.3: Constant hazard rate - źródło: [22]

Jak widać na powyższym rysunku, współczynnik hazardu pozostaje wartością stałą przez cały okres badania. Czas do wystąpienia zdarzenia opisujemy funkcją hazardu rozkładu wykładniczego ($h(t) = \lambda$). Bez względu na wartość t, $h(t)$ pozostaje niezmienna. W rzeczywistości stałą funkcję hazardu używa się rzadko, najczęściej na krótkich przedziałach czasowych.

2. *Increasing/Decreasing failure rate* - monotnicznie rosnące i zmniejszające się zagrożenie

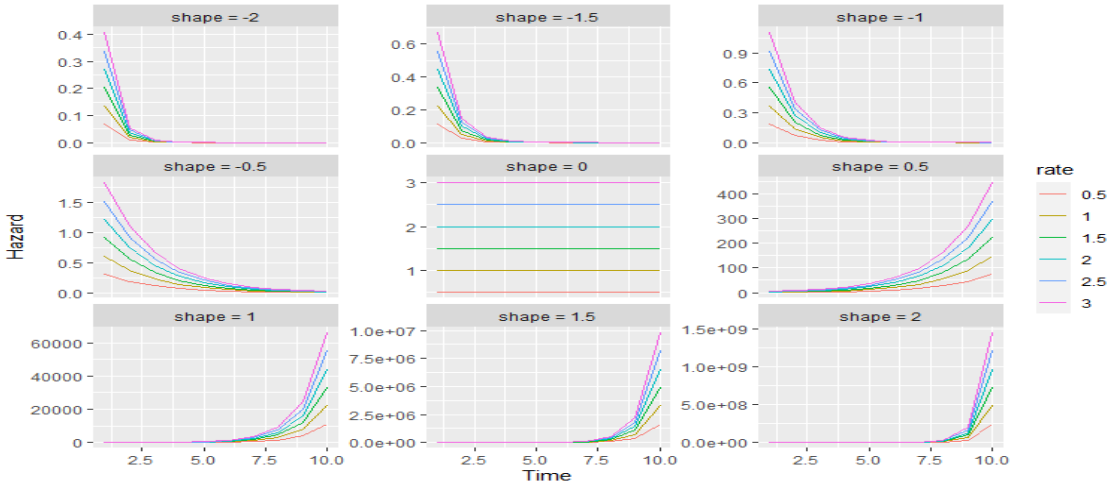


Rysunek 1.4: IFR/DFR - źródło: [22]

Rozkład czasu życia ma rosnącą intensywność awarii (*increasing failure rate - IFR*), jeżeli jego funkcja hazardu jest niemalejąca na przedziale $[0, \infty)$. Przykładem może być rozkład Weibulla $\mathrm{We}(\alpha, \beta)$, gdy $\alpha \geq 1$ oraz β - dowolne.
Rozkład czasu życia ma malejącą intensywność awarii (*decreasing failure rate - DFR*), jeżeli jego funkcja hazardu jest nierosnąca na przedziale $[0, \infty)$. Ponownie za przykład możemy podać rozkład Weibulla $\mathrm{We}(\alpha, \beta)$, tym razem dla $\alpha \in (0, 1]$ oraz dowolnego β.

Zauważmy, że dla $\alpha = 1$, rozkład Weibulla jest szczególnym przypadkiem rozkładu wykładniczego ze stałą funkcją hazardy. Zawiera się on zatem jednocześnie w *IFR* i *DFR*.

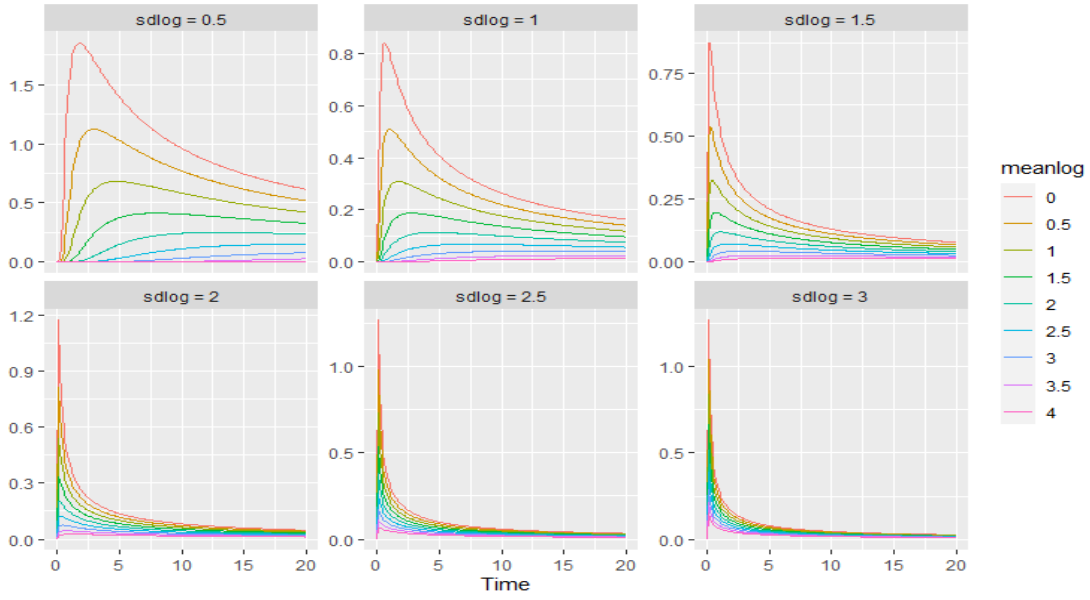
Rozkład Gompertza, określony przez parametr kształtu a i parametr b, także może być przykładem dla funkcji hazardu monotonicznie rosnącej, czy też malejącej.



Rysunek 1.5: Rozkład Gompertza - źródło: [22]

Widzimy, że $h(t)$ rośnie dla $a > 0$, jest stałe dla $a = 0$ oraz maleje dla $a < 0$. Można zauważyć, że dla $a = 0$, rozkład Gompetza jest równoważny rozkładowi wykładniczemu ze stałym parametrem b.

3. *Bathtub - shaped failure rate/Upside - down bathtub - shaped failure rate* - krzywa wannowa oraz odwrócona krzywa wannowa



Rysunek 1.6: Bathtub - shaped - źródło: [22]

Powyższe wykresy przedstawiają rozkład log - normalny, z różnymi parametrami, o funkcji gęstości:

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{(\ln t - \mu)^2}{2\sigma^2}} \tag{1.4}$$

parametryzowany przez średnią (μ) oraz odchylenie standardowe (σ) czasu przeżycia na skali logarytmicznej.

Krzywa wannowa opisuje szczególną postać funkcji hazardu składającą się z trzech części:

- pierwszą częścią jest rosnący wskaźnik awaryjności (*failure rate*)
- drugą częścią jest stały współczynnik awaryjności (*constance failure rate*)
- trzecią częścią nazywamy malejący wskaźnik awaryjności

Przykładem rozkładu o wannowym (*Bathtub-shaped failure rate*) kształcie intensywności może być także rozkład wprowadzony przez Chena BS(α, β) (gdzie $\alpha > 0$, $\beta > 0$ są parametrami kształtu) o funkcji gęstości:

$$f(t; \alpha, \beta) = \alpha\beta \exp(\alpha) t^{\beta-1} \exp(t^\beta) exp(-\alpha \exp(t^\beta)) \mathbf{1}_{(0,\infty)}(t) \tag{1.5}$$

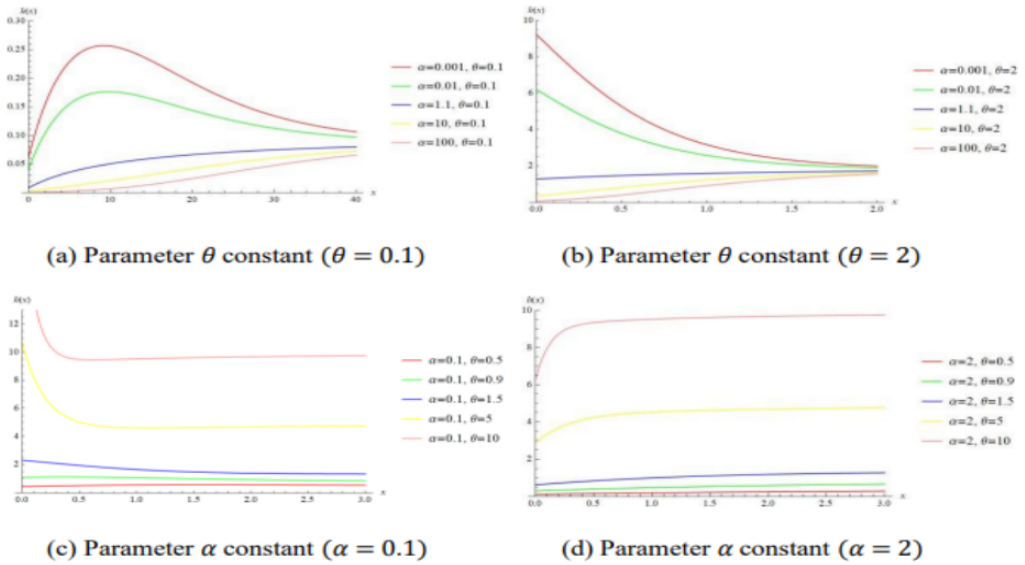
Dystrybuanta tego rozkładu ma postać:

$$F(t;\alpha,\beta) = 1 - \exp(\alpha(1 - \exp(t^{\beta}))) \tag{1.6}$$

Stąd otrzymujemy funkcję przeżycia:

$$S(t;\alpha,\beta) = 1 - F(t;\alpha,\beta) = \exp(\alpha(1 - \exp(t^{\beta}))) \tag{1.7}$$

Funkcję awaryjności o kształcie odwróconej krzywej wannowej właściwie można uzyskać z połączenia dwóch rosnących modeli funkcji IFR.



(a) Parameter θ constant ($\theta = 0.1$) (b) Parameter θ constant ($\theta = 2$)

(c) Parameter α constant ($\alpha = 0.1$) (d) Parameter α constant ($\alpha = 2$)

Rysunek 1.7: Upside - down bathtub - shaped - źródło: [22]

Cumulative hazard function to całkowanie współczynnika zagrożenia w zależności od czasu. Korzystając z tego, że:

$$h(t) = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)}$$

Otrzymujemy zależność funkcji skumulowanego hazardu od dystrybuanty.

$$\Lambda(t) = -ln[S(t)] = -ln[1 - F(t^{-})]$$

W poniższej tabelce porównamy funkcję gęstości, hazardu, dystrybuantę oraz funkcję przeżycia dla rozkładu wykładniczego E(λ), Weibulla We(α,β), Log normalnego i Gompertza.

	E(λ)	We(α,β)	LogN(μ,σ^2)	Rozkład Gompertza
f(t)	$\lambda e^{-\lambda t}$	$\frac{a}{b}\left(\frac{t}{b}\right)^{a-1} e^{-(t/b)^a}$	$\frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{(\ln t-\mu)^2}{2\sigma^2}}$	$be^{at}\exp\left[-\frac{b}{a}(e^{at}-1)\right]$
h(t)	λ	$\frac{a}{b}\left(\frac{t}{b}\right)^{a-1}$	$f(t)/S(t)$	be^{at}
F(t)	$1 - e^{-\lambda t}$	$1 - e^{-(t/b)^a}$	$\Phi\left(\frac{\ln t-\mu}{\sigma}\right)$	$1 - \exp\left[-\frac{b}{a}(e^{at}-1)\right]$
S(t)	$e^{-\lambda t}$	$e^{-(t/b)^a}$	$1 - \Phi\left(\frac{\ln t-\mu}{\sigma}\right)$	$\exp\left[-\frac{b}{a}(e^{at}-1)\right]$

Tabela 1.1: Zestawienie charakterystyk rozkładów - opracowanie własne

Rozdział 2

MODELE I METODY ANALIZY PRZEŻYCIA

2.1 Estymator Kaplana - Meiera

Analiza czasu przeżycia jest analizą czasu trwania. Pozwala ona na wyznaczenie związku czynników badanych z punktem końcowym (czyli zdarzeniem będącym przedmiotem badań) w sytuacjach, gdy informacja dotycząca wystąpienia punktu końcowego u wszystkich badanych, jest ograniczona.
Najstarszym sposobem do opisywania czasów przeżycia są tablice trwania życia (*life tables*). Składają się one z kilku funkcji względem wieku powiązanych ze sobą matematycznie oraz określają teoretyczny proces wymierania populacji w miarę jej starzenia się. Wartości tych funkcji oblicza się na podstawie liczby zgonów i liczby ludności badanej populacji według płci i wieku, zaobserwowanych w danym okresie. Narzędzie to wykorzystuje się w badaniu procesów demograficznych czy też w firmach zajmujących się ubezpieczeniami na życie. W wielu problemach estymacji niemożliwe jest wykonanie pełnych pomiarów na wszystkich elementach naszej próby. Na przykład w badaniach medycznych mających na celu określenie czasów przeżycia po operacji, kontakt z niektórymi osobami może zostać utracony przed śmiercią, zaś inne osoby odejdą z przyczyn, które należy wykluczyć z naszych rozważań.

Zatem jaką metodą mamy obliczać funkcje przeżycia w przypadku długich (paroletnich) badań, z dużą liczbą ilości danych cenzurowanych?
Kaplan i Meier [25] odpowiedzieli na zadane powyżej pytanie. Zaproponowali nową metodę estymacji funkcji przeżycia, która bierze pod uwagę obserwacje ucięte (zwłaszcza dane cenzurowanie prawostronnie). Co ciekawe, raport ten, od roku wydania został zacytowany prawie 61000 razy stając się tym samym najczęściej przytaczaną publikacją statystyczną w literaturze naukowej.

Estymator Kaplana - Meiera (KM) w badaniach medycznych może zostać użyty na przykład do określenia frakcji pacjentów, którzy przeżyją określony czas po operacji. Ekonomista może szacować czas, przez jaki ludzie pozostają bezrobotni po stracie pracy, czy na przykład inżynier może mierzyć czas do awarii komputera. O popularności tej metody może także świadczyć fakt implementacji w kilku językach programowania.

W szczególności możemy wymienić:

- `SAS` przez całą procedurę `PROC LIFETEST`
- `R` za pośrednictwem biblioteki `survival`
- `Stata` za pośrednictwem polecenia `sts`
- `Python` z pakietami `lifelines` i `scikit - survival`

Poniżej zaprezentujemy 3 sposoby implementacji estymatora Kaplana - Meiera w językach programowania `SAS`, `Stata` oraz `Python`. Pomijamy na razie język `R`, ponieważ w rozdziale 3.3 (*Analiza przeżycia na podstawie danych z biblioteki* ***`survival`*** *w programie* ***`R`***) znajduje się szczegółowa analiza przykładowych danych `lung` z biblioteki `survival`, w tym oczywiście, estymator KM.

2.1.1 Estymator Kaplana - Meiera w języku programowania `SAS`

Procedura `LIFETEST` oblicza nieparametryczne oszacowania funkcji rozkładu przeżycia. Możemy uzyskać limit produktu (*Kaplan - Meier*) bądź oszacowanie rozkładu życia (*life - table*). Funkcja `PROC LIFETEST` oblicza testy nieparametryczne w celu porównania krzywych przeżycia dwóch, lub więcej grup.

Rozpatrujemy dane[1] dotyczące 137 u pacjentów po przeszczepieniu szpiku kostnego przeprowadzonego przez *Klein i Moeschberger (*1997*)*, które są dostępne w bibliotece `Sashelp`. W momencie przeszczepu każdy z pacjentów klasyfikowany jest w jednej z trzech kategorii ryzyka.

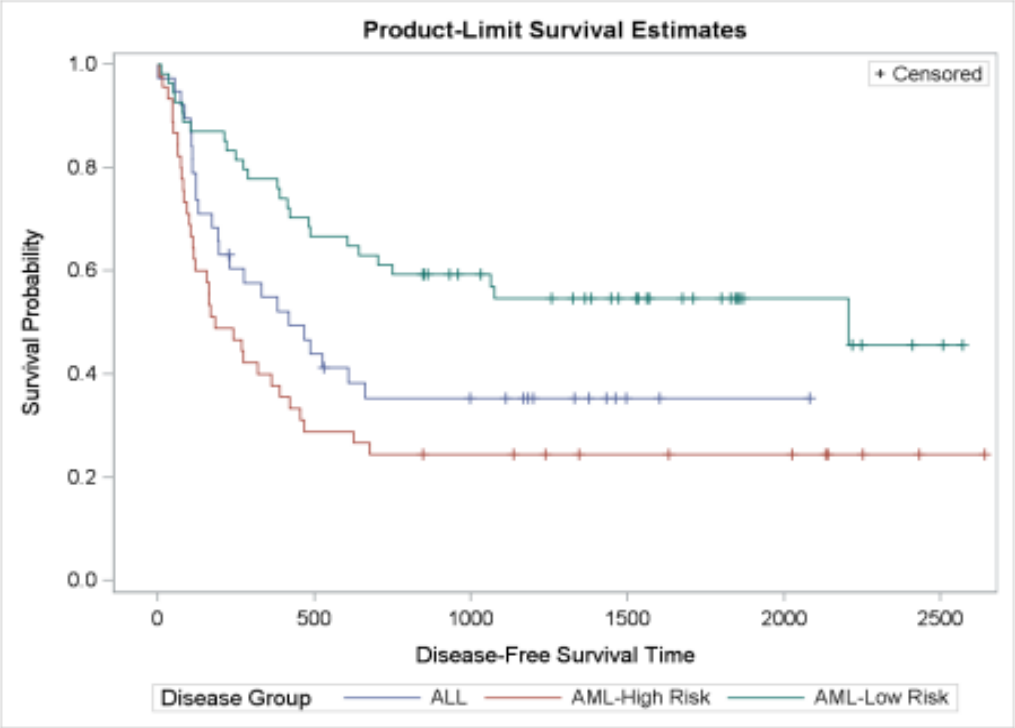
- ALL - ostra białaczka limfoblastyczna
- AML - Low Risk - ostrabiałaczka mielocytowa - niskie ryzyko
- AML - High Risk - ostrabiałaczka mielocytowa - wysokie ryzyko

Interesującym nas zdarzeniem jest przeżycie osoby, czyli czas w dniach do śmierci, nawrotu choroby czy też zakończenia badania. Zmienna *Group* reprezentuje kategorię ryzyka pacjenta, zaś zmienna T - czas przeżycia bez choroby. Zmienna *Status* jest wskaźnikiem cenzurowania (1 - oznacza, że nastąpiło zdarzenie, 0 - oznacza, że nastąpiło ocenzurowanie zmiennych)
Poniżej zaprezentujemy, w jaki sposób możemy uzyskać wykres estymatora Kaplana - Meiera funkcji przeżycia korzystając z `PROC LIFETEST`.

```
ods graphics on;

proc lifetest data=sashelp.BMT;
   time T * Status(0);
   strata Group;
run;
```

Rysunek 2.1: Kod źródłowy w programie `SAS` - źródło: [42]

[1]Dane dostępne między innymi w książce *Sashelp Data Sets*, ogólnodostępnej pod adresem https://support.sas.com/documentation/onlinedoc/stat/132/sashelp.pdf



Rysunek 2.2: Krzywa Kaplana - Meiera w programie `SAS` - źródło: [**9**]

Powyższy wykres składa się z trzech kolorowych funkcji krokowych (po jednej dla każdej z trzech grup ryzyka). Wykres pokazuje, że pacjenci w grupie AML - Low Risk cechują się dłuższymi czasami przeżycia bez choroby niż pacjenci z grup ALL bądź AML - High Risk.

2.1.2 Estymator Kaplana - Meiera w języku programowania `Stata`

Funkcja `sts` w programie `Stata` raportuje oraz tworzy zmienne zawierające szacowany czas przeżycia. Dla funkcji przeżycia `sts` produkuje oszacowania Kaplana - Meiera, lub, poprzez regresję Coxa, skorygowane oszacowania.

Poniżej zaprezentujemy, w jaki sposób możemy nakreślić krzywą estymatora KM. Skorzystamy w tym celu z danych *Heart transplant data*[2].

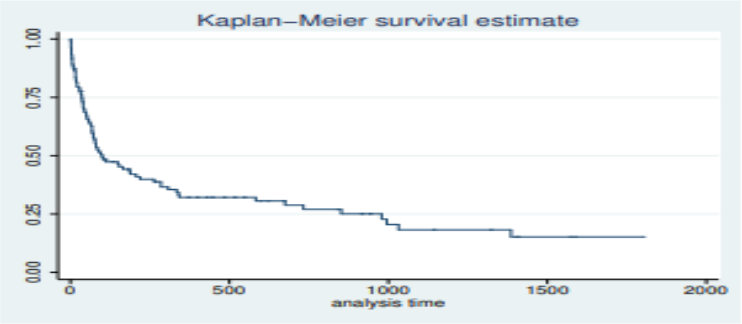
```
. use http://www.stata-press.com/data/r13/stan3
(Heart transplant data)
```

Rysunek 2.3: Kod źródłowy w programie `Stata` - źródło[3]

[2]Dane dostępne między innymi pod adresem https://www.stata-press.com/data/r13/stan3

[3]https://www.stata.com/support/faqs/graphics/gph/graphdocs/kaplan-meier-survival-function/index.html

```
. sts graph
```



Rysunek 2.4: Krzywa Kaplana - Meiera w programie `Stata` - źródło[4]

Na powyższym wykresie widzimy bardzo dużą śmiertelność na początku badania, która z czasem się stabilizuje.

2.1.3 Estymator Kaplana - Meiera w języku programowania `Python`

W celu zilustrowania estymatora KM w programie `Python` wykorzystamy dane *echocardiogram - UCI* zawierające dane[5] na temat przeżycia co najmniej rok pacjentów po zawale serca.

Po wczytaniu wszystkich odpowiednich bibliotek zabieramy się za napisanie kodu tworzącego krzywą Kaplana - Meiera.

```python
# Import required libraries :

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from lifelines import KaplanMeierFitter
```

Rysunek 2.5: Wczytywanie bibliotek w języku `Python` - źródło: [9]

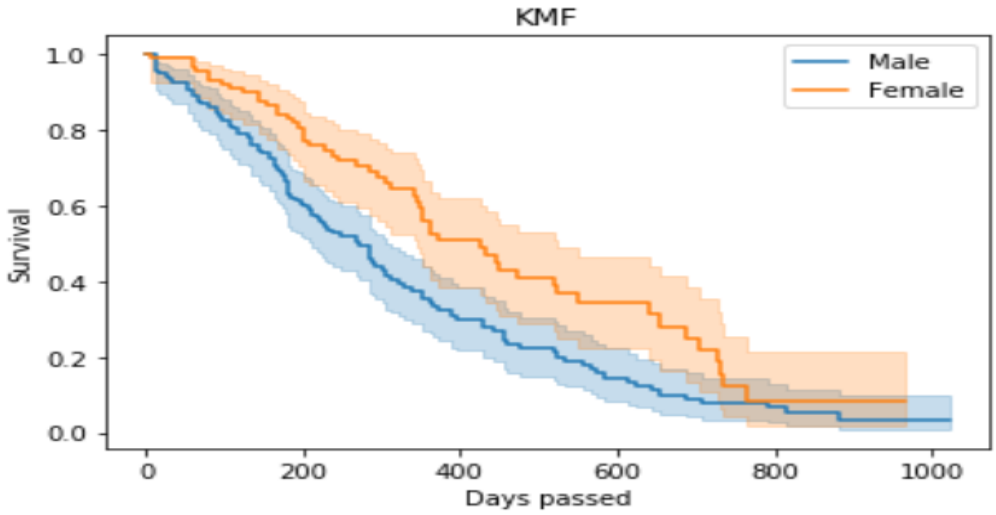
```python
#Read the dataset :

data = pd.read_csv("lung.csv")
data.head()
```

	Unnamed: 0	inst	time	status	age	sex	ph.ecog	ph.karno	pat.karno	meal.cal	wt.loss
0	1	3.0	306	2	74	1	1.0	90.0	100.0	1175.0	NaN
1	2	3.0	455	2	68	1	0.0	90.0	90.0	1225.0	15.0
2	3	3.0	1010	1	56	1	0.0	90.0	90.0	NaN	15.0
3	4	5.0	210	2	57	1	1.0	90.0	60.0	1150.0	11.0
4	5	1.0	883	2	60	1	0.0	100.0	90.0	NaN	0.0

Rysunek 2.6: Wczytywanie danych w języku `Python` - źródło: [9]

[4]https://www.stata.com/support/faqs/graphics/gph/graphdocs/kaplan-meier-survival-function/index.html

[5]Dane dostępne pod adresem https://www.kaggle.com/datasets/loganalive/echocardiogram-uci



Rysunek 2.7: Krzywa Kaplana - Meiera w programie `Python` - źródło: [9]

Analizując wszystkie powyższe wykresy zauważamy kilka cech krzywej Kaplana - Meiera.

- Jest to funkcja składająca się z szeregu poziomych odcinków, schodzących coraz niżej, czyli tak zwana funkcja schodkowa (co dość dokładnie widać na wykresie 2.7).
- Czas przeżycia osobnika jest reprezentowany przez długości linii poziomych wzdłuż osi X kolejnych czasów.
- Coraz większa próba statystyczna spowoduje powstawanie coraz większej liczby coraz krótszych odcinków, w granicy dążąc do prawdziwej funkcji przeżycia (dla dużych prób rozkład estymatora KM zmierza do rozkładu normalnego).
- Pionowe kreski widoczne na powyższych wykresach są zaznaczonymi obserwacjami ocenzurowanymi.

Ponadto, istotną zaletą estymatora KM jest fakt, że nie tylko nie odrzuca on obserwacji uciętych poprzez przypisanie im etykiety braku danych, ale bierze je pod uwagę przy obliczaniu funkcji przeżycia, jest fakt, że na uzyskiwane oceny nie wpływa grupowanie danych w przedziałach czasowych.

PRZYKŁAD
Niech $S(t)$ będzie funkcją przeżycia, oznaczającą prawdopodobieństwo, że element populacji przeżyje co najmniej t. Uporządkujmy teraz N - elementową próbę z tej populacji według czasu przeżycia: $t_1 \leq t_2 \leq \ldots \leq t_N$. Z każdym t_i związana jest pewna liczba n_i tych, o których wiemy, że dożyli do tego momentu oraz d_i oznaczająca liczbę śmierci w momencie t_i. Zauważmy, że odległości pomiędzy kolejnymi momentami t_i zwykle nie będą stałe. Na przykład jeśli rozpatrujemy 20 przypadków, ze śmiercią w dniu 5, utratą kontaktu (czyli obserwacją ocenzurowaną) w dniu 10 i kolejną śmiercią w dniu 15, to wówczas:

$$t_1 = 5,\ n_1 = 20,\ d_1 = 1$$
$$t_2 = 15,\ n_2 = 18,\ d_2 = 1$$

Estymator Kaplana - Meiera jest nieparametrycznym estymatorem największej wiarygodności funkcji S(t). Jest to iloczyn postaci:

$$\widehat{S}(t) = \Pi_{t_i<t}\frac{n_i - d_i}{n_i} \tag{2.1}$$

Bądź alternatywnie używa się wzoru:

$$\widehat{S}(t) = \Pi_{t_i\leq t}\frac{n_i - d_i}{n_i} \tag{2.2}$$

Powyższe definicje różnią się jedynie warunkiem na czas. We wzorze (2.1) estymator jest funkcją lewostronnie ciągłą, zaś we wzorze (2.2) - prawostronnie ciągłą.

Ustalmy, że F jest dystrybuantą czasu życia i $T > 0$ będzie takie, że $F(T) < 1$. Wówczas, dla każdego $t \in (0, T)$ estymator Kaplana - Meiera $\widehat{S}_n(t)$ w punkcie t zbiega według prawdopodobieństwa do $S(t)$ dla $n \to \infty$ (zostało to udowodnione w [25]). Ponadto, Breslow i Crowley [7] pokazali, że dla niezależnych dystrybuant F i G (odpowiednio czasu życia i czasu cenzurowania) ciągłych na $[0, T]$ i $F(T) < 1$, dla każdego $t \in (0, T)$, estymator $\widehat{S}_n(t)$ w punkcie t jest asymptotycznie normalny o wartości oczekiwanej $S(t)$ i wariancji zadanej wzorem:

$$\sigma_n^2 = \frac{S^2(t)}{n}\int_0^t \frac{dF_u(u)}{(1-\Lambda(u))^2} \tag{2.3}$$

Oznacza to, że wyrażenie:

$$\frac{\widehat{S}_n(t) - S(t)}{\sigma_n(t)} \tag{2.4}$$

Dąży według rozkładu do standardowego rozkładu normalnego, gdy rozmiar próby n dąży do nieskończoności. Wynika stąd dalej, że dla dostatecznie dużego n zachodzi:

$$P\left(\frac{\widehat{S}_n(t) - S(t)}{\sigma_n(t)} \leq x\right) \approx \Phi(x) \tag{2.5}$$

Gdzie $\Phi(x)$ oznacza dystrybuantę standardowego rozkładu normalnego.

2.2 Metody estymacji funkcji przeżycia

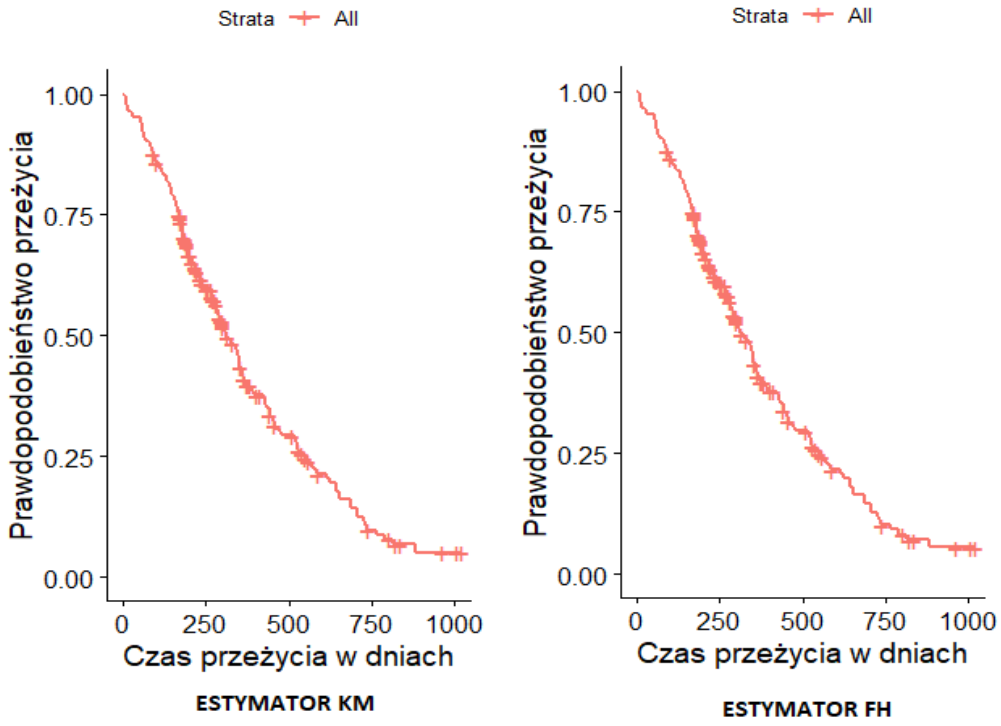
2.2.1 Nieparametryczne

Estymator Kaplana - Meiera został omówiony w rozdziale 2.1 *Estymator Kaplana - Meiera*, zatem zajmiemy się teraz przestudiowaniem estymatora Flemingtona - Harringtona. Flemington i Harrington [18] zaproponowali nową klasę testów, gdzie jako wagi przyjęli iloczyn funkcji przeżycia i dystrybuanty rozkładu, ważony odpowiednimi potęgami. Jest on w następującej postaci:

$$W(t) = W_{pq}(t) = \left(\widehat{S}(t)\right)^p\left(1 - \widehat{S}(t)\right)^q \tag{2.6}$$

Gdzie $p \geq 0$, $q \geq 0$ są parametrami, których specyfikacja jest dowolna. W przypadku, gdy $p < q$ większą wagę przypisujemy podmiotom mającym dłuższe czasy trwania, zaś dla $p > q$ jest odwrotnie. Ponadto $\widehat{S}(t)$ oznacza estymator Kaplana - Meiera funkcji przeżycia.

Estymator Fleminga - Harringtona powinien dawać bardzo porównywalne oszacowania funkcji przeżycia z estymatorem Kaplana - Meiera. Poniżej przedstawione zostały dwa nieparametryczne estymatory funkcji przeżycia.



Rysunek 2.8: Porównanie wykresów estymatora KM oraz FH - opracowanie własne

Przyglądając się powyższym wykresom, nie widzimy żadnych znaczących różnic, co pozwala sądzić, że otrzymaliśmy oczekiwane przez nas założenie - estymator Fleminga - Harringtona jest bardzo zbliżony do estymatora Kaplana - Meiera.

2.2.2 Parametryczne

W powyższym podrozdziale rozpatrywaliśmy problemy statystyczne w przypadku gdy nie zakładaliśmy postaci rozkładu obserwowalnej zmiennej losowej. Może się jednak zdarzyć, że będziemy posiadać dodatkowe informacje (zwane charakterystykami jednostki) na temat badanych zjawisk, co pozwoli nam na prognozowanie przy użyciu dopasowanego modelu.

Rozkład	$S(t)$	$h(t)$
Wykładniczy $E(\lambda)$	$e^{-\lambda t}$	λ
Weibulla $We(\alpha, \beta)$	$e^{-(t/\beta)^{\alpha}}$	$\frac{\alpha}{\beta}\left(\frac{t}{\beta}\right)^{\alpha-1}$
Log - logistyczny $LL(\alpha, \lambda)$	$\frac{t/\lambda)^{-\alpha}}{(1+(t/\lambda)^{-\alpha})}$	$\frac{\alpha}{t(1+(t/\lambda)^{-\alpha}}$

Tabela 2.1: Funkcje przeżycia oraz hazardu wybranych estymatorów parametrycznych - opracowanie własne

Powyższa tabela zawiera informacje na temat funkcji przeżycia oraz funkcji hazardu trzech najpowszechniejszych dystrybucji używanych do przetrwania.

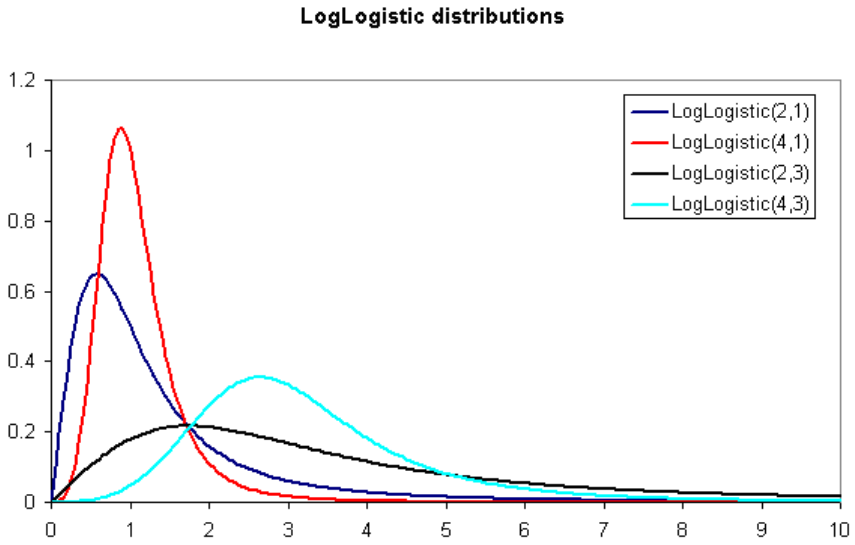
Rozkłady wykładniczy $E(\lambda)$ oraz Weibulla $We(\alpha, \beta)$ zostały omówione w rozdziale 1.3, zatem teraz zajmiemy się rozkładem log - logistycznym (LL). Ma on dość elastyczną formę funkcjonalną - współczynnik ryzyka początkowo wzrasta, a następnie maleje. Czasami może mieć on kształt garbu. Rozkład log - logistyczny może zostać użyty niekiedy jako odpowiednik rozkładu Weibulla, jednak jest on połączeniem rozkładu Gompertza oraz rozkładu gamma. Gęstość rozkładu LL zadana jest wzorem:

$$f(t;\alpha,\lambda) = \frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1}\left(1+\left(\frac{t}{\lambda}\right)^{\alpha}\right)^{-2} \tag{2.7}$$

Oraz dystrybuantę określamy jako:

$$F(t;\alpha,\lambda) = \left(1+\left(\frac{t}{\lambda}\right)^{-\alpha}\right)^{-1} \tag{2.8}$$

Poniżej zobrazowany został przykładowy wykres rozkładu $LL(\alpha, \lambda)$ dla różnych parametrów α i β. Widzimy, że krzywe są w kształcie odwróconej krzywej wannowej.



Rysunek 2.9: Wykres rozkładu log - logistycznego dla różnych parametrów - źródło[6]

2.3 Testy statystyczne w analizie przeżycia

Jednym z podstawowych zagadnień w analizie przeżycia jest testowanie, czy dwie wybrane próby mają te same funkcje przeżycia. W tej części przyjrzymy się dwóm nieparametrycznym testom statystycznym.

Test Kołmogorowa - Smirnowa opiera się na maksymalnej różnicy między empirycznym a hipotetycznym rozkładem symulowanym. Główną wadą tego testu jest fakt, że może on zostać użyty jedynie w przypadku braku danych cenzurowanych, co w analizie przeżycia jest dość rzadko spotykanym przypadkiem. Test Kołmogorowa - Smirnowa ocenia zgodność

[6]https://www.vosesoftware.com/riskwiki/Loglogisticdistribution.php

rozkładu analizowanych zmiennych z rozkładem normalnym. To znaczy, przyjmujemy hipotezę zerową (H_0) wskazującą na rozkład zbliżony do rozkładu normalnego przeciwko hipotezie alternatywnej (H_1) mówiącej o nierówności rozkładu zmiennych losowych. Przy ustalonym poziomie istotności (na przykład $\alpha = 0.05$), dla wartości $p - value > \alpha$ otrzymujemy potwierdzenie spełnienia założenia o rozkładzie normalnym, czyli przyjmujemy hipotezę zerową. W przeciwnym wypadku (dla $p - value < \alpha$) odrzucamy hipotezę zerową i przyjmujemy alternatywną o nierówności badanych rozkładów. Test Kołmogorowa - Smirnowa zaimplementowany jest w języku programowania `R` jako funkcja `ks.test`.

PRZYKŁAD
Rozważmy dwa przypadki. Niech X pochodzi z rozkładu normalnego, Y z rozkładu jednostajnego, Z oraz K z rozkładu logistycznego. Korzystając z funkcji `ks.test` wykonamy dwa testy statystyczne.

- H_{01} mówiąca o równości rozkładów X i Y przeciwko H_{11} mówiącej o nierówności tych rozkładów
- H_{02} mówiąca o równości rozkładów Z i K przeciwko H_{12} mówiącej o nierówności tych rozkładów

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  X and Y
## D = 0.36, p-value = 0.01184
## alternative hypothesis: two-sided
```

Rysunek 2.10: Test Kołmogorowa - Smirnowa dla zmiennych X i Y - opracowanie własne

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  Z and K
## D = 0.20667, p-value = 0.3568
## alternative hypothesis: two-sided
```

Rysunek 2.11: Test Kołmogorowa - Smirnowa dla zmiennych Z i K - opracowanie własne

Jak wynika z rysunku (2.10), $p - value = 0.01184 < 0.05 = \alpha$, zatem odrzucamy hipotezę zerową i przyjmujemy hipotezę alternatywną. Zaś dla zmiennych Z oraz K (2.11) otrzymaliśmy $p - value = 0.3568 > 0.05 = \alpha$, więc przyjmujemy hipotezę zerową.

Popularniejszym testem jest test log - rank, który może zostać użyty także w przypadku wystąpienia wartości cenzurowanych. Po raz pierwszy został on opisany przez Mantela [31] i nazywany jest również testem Mantela - Coxa. Statystyka tego testu porównuje oszacowania funkcji hazardu dwóch grup w określonym czasie zdarzenia. Zasada stojąca za testem log - rank dla porównania dwóch tablic trwania życia jest prosta - jeżeli nie było różnic między grupami, całkowita liczba zgonów występujących w dowolnym czasie

powinna zostać podzielona między dwie grupy w tym czasie. Przykładowo, jeśli liczba zagrożonych w pierwszej i drugiej grupie w siódmym miesiącu wynosiła odpowiednio 60 i 40, a 20 zgonów miało miejsce w tym miesiącu, oczekujemy, że $20 \cdot \frac{60}{60+40} = 12$ zgonów miało miejsce w pierwszej grupie, oraz $20 \cdot \frac{40}{60+40} = 8$ zgonów miało miejsce w drugiej grupie. Hipoteza zerowa dla tego testu zakłada brak różnic w rozkładzie przeżycia w obu grupach. Sprawdzenie jej wymaga spojrzenia na różnice w liczbie obserwacji w każdej grupie, które doświadczyły zadanego zdarzenia w poszczególnych przedziałach względem ich spodziewanej liczby. Test log - rank zaimplementowany jest w pakiecie `R` jako funkcja `survdiff`.

PRZYKŁAD
Korzystając z testu log - rank na podstawie danych `lung` oraz zmiennej sex zbadamy hipotezę zerową H_0 mówiącą o braku różnić w przeżyciu dla płci, przeciwko hipotezie alternatywnej H_1 mówiącej o występujących różnicach dla przeżycia według zmiennej płci.

```
## Call:
## survdiff(formula = Surv(time, status) ~ sex, data = lung)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=1 138      112     91.6      4.55      10.3
## sex=2  90       53     73.4      5.68      10.3
##
##  Chisq= 10.3  on 1 degrees of freedom, p= 0.001
```

Rysunek 2.12: Test log - rank dla zmiennej sex - opracowanie własne

Statystyka testu Chi - kwadrat wynosi 10.3 z 1 stopniem swobody, zaś odpowiadająca jej wartości $p - value$ wynosi 0.001, co jest mniejsze od poziomu istotności $\alpha = 0.05$. W takiej sytuacji odrzucamy hipotezę zerową. Nasuwa się oczywisty wniosek, że istnieje statystycznie istotna różnica w przeżywalności między dwiema badanymi grupami.

2.4 Nieparametryczny model Coxa

Jak już określiliśmy w rozdziale 1.3, analiza przeżycia jest gałęzią statystyki obejmującą metody badania procesów, w których interesuje nas czas, jaki upłynie do wystąpienia pewnego zdarzenia. Głównymi pytaniami, na jakie analiza przeżycia może nam udzielić odpowiedzi, to:

- Jaki odsetek populacji przetrwa założony okres czasu?
- Ile czynników może wpływać na długość okresu przetrwania?
- Jak długo będą żyć osoby, które przetrwają zdefiniowany okres naszego badania?

Jedną z typowych metod analizy przeżycia są modele regresji, w tym model proporcjonalnych hazardów Coxa, dzięki któremu jesteśmy w stanie wyizolować poszczególne elementy (zmienne niezależne/objaśniające) mające wpływ na prognozowanie *failure time*

Podczas modelowania funkcji przeżycia wprowadza się nową zmienną - funkcję hazardu, zadaną wzorem (1.2). Najczęściej zakłada się, że na wartość hazardu mają wpływ różne

zmienne objaśniające. Cox [8] wprowadził pojęcie hazardu proporcjonalnego. Zakłada on, że nieznany hazard jest funkcją zmiennych niezależnych określonych wzorem:

$$r(x, \beta) = e^{x\beta} \tag{2.9}$$

Oraz funkcję hazardu postaci:

$$h(t, x, \beta) = h_0(t)e^{x\beta} \tag{2.10}$$

Gdzie $h_0(t)$ jest poziomem hazardu bazowego, $\beta = (\beta_1, \ldots, \beta_n)$ jest wektorem współczynników regresji, zaś x definiujemy jaki wektor wartości zmiennych objaśniających.

Model Coxa nazywany jest modelem hazardu proporcjonalnego, ponieważ zakładamy, że stosunek ryzyk dwóch przypadków nie zależy od czasu, to znaczy, jest postaci:

$$\frac{h(t, x_1, \beta)}{h(t, x_2, \beta)} = \frac{h_0(t)\exp\{x_1\beta\}}{h_0(t)\exp\{x_2\beta\}} = \exp\{\beta(x_1 - x_2)\}$$

Stosunek tych dwóch funkcji określamy jako współczynnik hazardu (*hazard ratio*) i interpretujemy jako współczynnik ryzyka względnego. Krótko zapisujemy go w postaci:

$$HR(t, x_1, x_2, \beta) = \exp\{\beta(x_1 - x_2)\} \tag{2.11}$$

PRZYKŁAD
Dla zmiennej niezależnej takiej jak kolor oczu, z wartościami $x_1 = 1$ dla oczu niebieskich, oraz $x_2 = 0$ dla oczu brązowych, współczynnik hazardu wynosi:

$$HR(t, 1, 0, \beta) = \exp\{\beta(1 - 0)\} = e^{\beta}$$

Dla $\beta = \ln(3)$ otrzymujemy, że tempo umieralności osób niebieskookich jest trzykrotnie wyższe niż tempo umieralności osób o brązowych oczach.
Funkcję przeżycia można wyrazić za pomocą wzoru:

$$S(t, x, \beta) = e^{-\Lambda(t,x,\beta)} \tag{2.12}$$

Gdzie $\Lambda(t, x, \beta)$ jest funkcją skumulowanego hazardu w chwili t dla niezależnej zmiennej x. W modelu Coxa funkcja skumulowanego hazardu jest postaci:

$$\Lambda(t, x, \beta) = \int_0^t h(u, x, \beta)\ du = r(x, \beta)\int_0^t h_0(u)\ du = r(x, \beta)H_0(t) \tag{2.13}$$

Czyli funkcję przeżycia zapisujemy jako:

$$S(t, x, \beta) = e^{-r(x,\beta)H_0(t)} = [e^{-H_0(t)}]^{r(x,\beta)} = [S_0(t)]^{r(x,\beta)} \tag{2.14}$$

Korzystając ze wzoru (2.9) otrzymujemy ostateczną wersję funkcji przeżycia w modelu Coxa przyjmująca wartości między 0 a 1.

$$S(t, x, \beta) = [S_0(t)]^{exp\{x\beta\}} \tag{2.15}$$

Jedną z głównych zalet modelu Coxa jest fakt, że można go stosować, gdy zmienna zależna nie ma rozkładu normalnego oraz gdy mamy do czynienia z niepełnymi danymi.

Hipotezą zerową w teście Coxa, podobnie jak w teście log - rank, określamy, że funkcje przeżycia w dwóch badanych grupach nie różnią się istotnie. W języku programowania `R`,

model Coxa można znaleźć w bibliotece `survival` pod nazwą `coxph`.

PRZYKŁAD
Korzystając z nieparametrycznego testu Coxa na podstawie danych lung oraz zmiennej sex zbadamy hipotezę zerową H_0 mówiącą o braku różnić w przeżyciu dla płci, przeciwko hipotezie alternatywnej H_1 mówiącej o występujących różnicach dla przeżycia według zmiennej płci.

```
## Call:
## coxph(formula = Surv(time, status) ~ sex, data = lung)
##
##        coef exp(coef) se(coef)      z       p
## sex -0.5310    0.5880   0.1672 -3.176 0.00149
##
## Likelihood ratio test=10.63  on 1 df, p=0.001111
## n= 228, number of events= 165
```

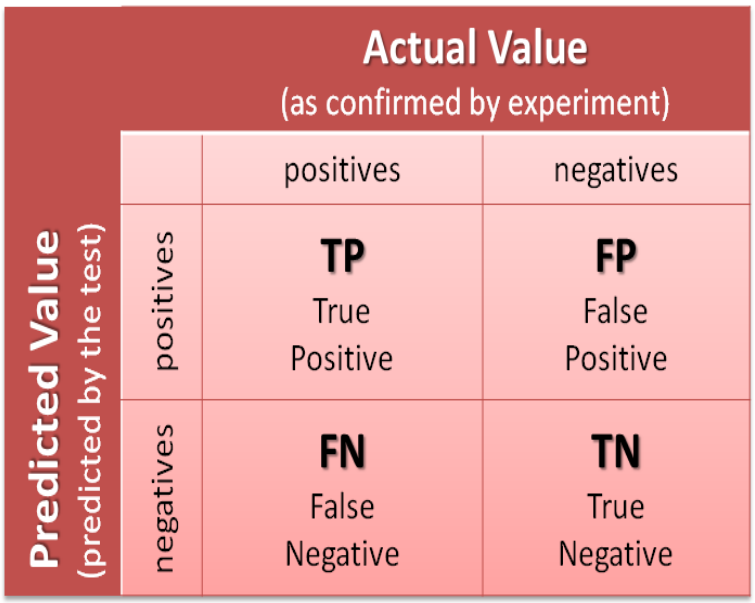
Rysunek 2.13: Nieparametryczny model Coxa - opracowanie własne

Zauważamy, że wartość $p - value = 0.001111 < 0.05 = \alpha$, więc odrzucamy hipotezę zerową i przyjmujemy alternatywną. Otrzymaliśmy w ten sposób identyczne wnioski, jak dla testu log - rank opisanego w rozdziale 2.3 (rysunek (2.12)).

2.5 Krzywa ROC

Krzywa ROC (*ang. Receiver Operating Characteristic*), czyli krzywa charakterystyki operatora odbiornika, jest wykresem graficznym używanym do pokazania możliwości diagnostycznych klasyfikatorów binarnych. Termin *charakterystyka pracy odbiornika* pochodzi od testów zdolności operatorów radarów z okresu II Wojny Światowej do określenia, czy plamka na ekranie radaru reprezentuje obiekt (sygnał), czy też zwykły hałas. Krzywe ROC zostały pierwotnie opracowane przez Brytyjczyków jako część systemu radarowego *Chain Home* (jest to kryptonim pierścienia przybrzeżnych stacji radarowych wczesnego ostrzegania). Tak jak kiedyś, angielscy żołnierze odszyfrowywali znaki na ekranie jako niemieckie bombowce czy przyjazne samoloty, tak dzisiaj radiolodzy stają przed zadaniem zidentyfikowania nieprawidłowej tkanki na skomplikowanym tle.

Analiza ROC opiera się na macierzach o wymiarach $N \times N$ nazywanych macierzami błędu (*ang. confusion matrix*). Używane są do wydajności modelu klasyfikacji, gdzie N oznacza liczbę klas docelowych. Macierz porównuje rzeczywiste wartości docelowe z przewidywanymi przed model. Daje nam to całościowy obraz tego, jak dobrze działa nasz model klasyfikacji i jakie rodzaje błędów popełnia.
W przypadku problemu klasyfikacji binarnej otrzymujemy macierz 2×2, która wygląda następująco:



Rysunek 2.14: Macierz błędów - źródło[7]

Kolumny macierzy reprezentują rzeczywiste wartości zmiennej docelowej, zaś wiersze - przewidywane wartości. Wyjaśnimy teraz kolejno pojęcia TP, FP, TN, FN na przykładzie testów medycznych.

- `TP` (*true positive*) ludzie chorzy poprawnie zdiagnozowani jako chorzy
- `FP` (*false positive*) ludzie zdrowi błędnie zdiagnozowani jako chorzy
- `TN` (*true negative*) ludzie zdrowi poprawnie zdiagnozowani jako ludzie zdrowi
- `FN` (*false negative*) ludzie chorzy błędnie zdiagnozowani jako zdrowi

Z macierzy pomyłek można wyliczyć wiele wskaźników dla klasyfikatora binarnego, takich jak:

- Dokładność (*Accuracy*) pozwala nam ocenić jakość klasyfikacji testu. Daje nam informacje na temat tego, jaka część testów, ze wszystkich zaklasyfikowanych, została oceniona poprawnie. Im wyższa wartość dokładności, tym lepiej. `ACC` $= 1$ oznacza idealnie dopasowanie i brak pomyłki ani razu.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.16}$$

- Prawdziwie pozytywna wartość (*True Positive Rate*) zwana inaczej czułością (*sensivity*). Mówi nam o tym, jaki jest udział prawidłowo prognozowanych przypadków pozytywnych wśród wszystkich przypadków pozytywnych. Wartość ta powinna być jak najbliższa 1.

$$TPR = \frac{TP}{TP + FN} \tag{2.17}$$

[7]Macierz błędów dostępna pod adresem https://mathspace.pl/matematyka/confusion-matrix-macierz-bledu-tablica-pomylek-czyli-ocena-jakosci-klasyfikacji-czesc-1/

- Fałszywie pozytywna wartość (*False Positive Rate*) określa jaki jest udział fałszywie pozytywnych przypadków wśród wszystkich negatywnych przypadków. Wartość jego powinna być jak najbliższa 0.

$$FPR = \frac{FP}{TN + FP} \tag{2.18}$$

- Prawdziwie negatywna wartość (*True Negative Rate*) inaczej swoistość (*specifity*) mierzy, jak dużo ze wszystkich negatywnych przypadków zostało rzeczywiście zaklasyfikowanych do tej kategorii.

$$TNR = \frac{TN}{TN + FP} \tag{2.19}$$

- Fałszywie negatywna wartość (*False Negative Rate* jest wynikiem wskazującym, że dany warunek istnieje, gdy go nie ma. Czyli na przykład test ciążowy wskazujący na ciążę kobiety, podczas gdy naprawdę nie jest w ciąży.

$$FNR = \frac{FN}{TP + FN} \tag{2.20}$$

- Precyzja przewidywania pozytywnego (*Positive Predictive Value*) daje nam informację, ile wśród przykładów ocenianych pozytywnie jest rzeczywiście pozytywnych. Wartość precyzji, podobnie jak dokładności i czułości, powinna przyjmować wartość jak najbliższą 1.

$$PPV = \frac{TP}{TP + FP} \tag{2.21}$$

- Precyzja przewidywania negatywnego (*Negative Predictive Value*), czyli miara precyzji wskazuje, z jaką pewnością możemy ufać przewidywaniom negatywnym.

$$NPV = \frac{TN}{TN + FN} \tag{2.22}$$

- Współczynnik korelacji Matthews (*ang. Matthews Correlation Coefficient*) przyjmujący wartości od -1 do 1. Dla $\texttt{MCC} = 1$ nasz model bardzo dobrze, wręcz idealnie klasyfikuje wszystko do prawidłowej kategorii, zaś dla $\texttt{MCC} = -1$ otrzymujemy informację, że wszystko zostało zaliczone do niepoprawnej kategorii.

$$MCC = \frac{TN \cdot TP - FP \cdot FN}{\sqrt{(TN + FN)(FP + TP)(TN + FP)(FN + TP)}} \tag{2.23}$$

PRZYKŁAD
Przeprowadzimy analizę danych rzeczywistych oraz danych prognozowanych dotyczących rozpoznawania kotów białych i czarnych, o łącznej ilości 205.

Rysunek 2.15: Przykładowa macierz błędów - opracowanie własne

Uzyskujemy następujące charakterystyki:

- TP = 101 - oznacza, że mamy faktycznie 101 kotów białych
- FP = 47 - oznacza, że ktoś ocenił, że 47 kotów białych jest kotami czarnymi
- FN = 32 - oznacza, że ktoś ocenił, że 32 czarne koty są kotami białymi
- TN = 25 - oznacza, że mamy faktycznie 25 kotów czarnych
- TP + TN = 126 - liczba poprawnych klasyfikacji
- FP + FN = 79 - liczba błędnych klasyfikacji

Następnie, korzystając z własnoręcznie napisanej funkcji w programie `R` wyznaczamy wartości wcześniej wymienionych wskaźników.

	Wartości
ACC	0.61
TPR	0.76
FPR	0.65
TNR	0.35
FNR	0.24
PPV	0.68
NPV	0.44
MCC	0.11

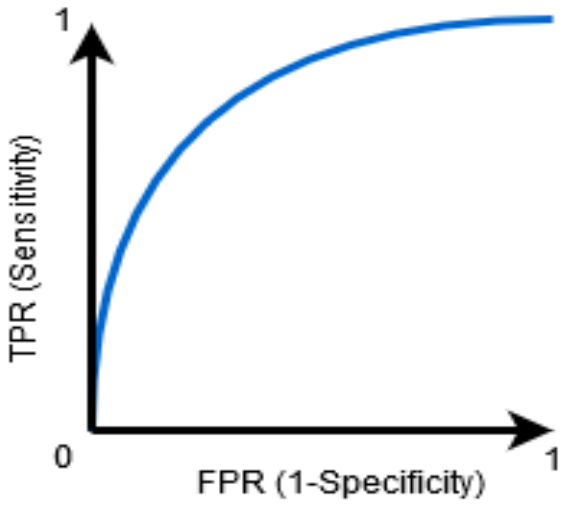
Tabela 2.2: Wyznaczone wartości wskaźników - opracowanie własne

Analizując niektóre powyższe wskaźniki, otrzymujemy, że jakoś naszej klasyfikacji (`ACC`) możemy określić na 61%, ufać przewidywaniom pozytywnym w 68% (`PPV`) a negatywnym - 44% (`NPV`). Klasa prawdziwie pozytywna została w 76% pokryta przewidywaniem pozytywnym (`TPR`), zaś klasa prawdziwie negatywna - 35% przewidywaniem negatywnym (`TNR`). Wskaźnik korelacji Matthewsa (`MCC`) wynosi 0.11, wskazuje na średnie przewidywanie losowe.

Krzywa ROC jest wykresem obrazującym wydajność modelu klasyfikacyjnego na wszystkich progach klasyfikacyjnych. Do jej konstrukcji wykorzystuje się dwa parametry:

- *True Positive Rate*
- *False Positive Rate*

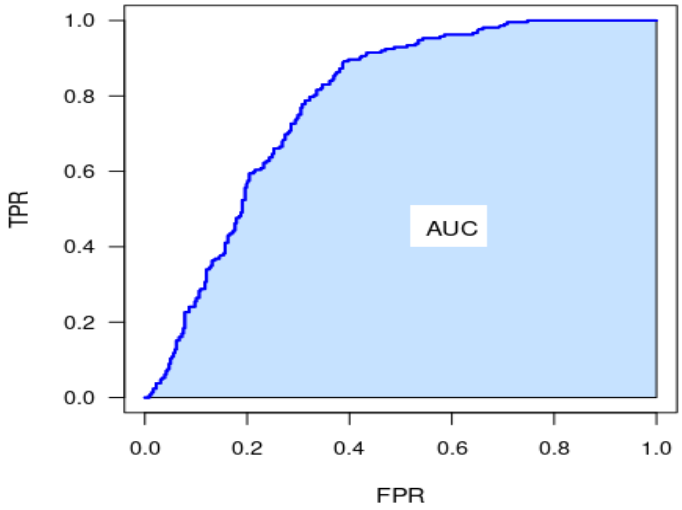
Krzywa ROC jest graficzną prezentacją wielkości separacji między rozkładem przypadków a kontrolnych markerów. Gdy pomiary przypadków oraz pomiary kontrolne nie pokrywają się, krzywa ROC przyjmuje wartość 1 (idealnie prawdziwie pozytywny wskaźnik *true positive*) dla każdego fałszywie pozytywnego wskaźnika (*false positive*) większego od 0. W przeciwnym wypadku krzywa ROC jest linią nachyloną do osi Ox pod kątem 45° i wskazuje, że marker jest złym wskaźnikiem do oddzielenia przypadków od obserwacji kontrolnych. Poniższy wykres przedstawia typowy wygląd krzywej ROC.



Rysunek 2.16: Wykres krzywej ROC - źródło[8]

Pole na wykresie, pod krzywą ROC, zwane AUC (*ang. Area Under the ROC Curve*) jest miarą zgodności między markerem a wskaźnikiem. Mierzy cały dwuwymiarowy obszar pod krzywą, od $(0, 0)$ do $(1, 1)$. Im wyższy AUC, tym lepsza wydajność modelu w rozróżnianiu klas pozytywnych i negatywnych.

[8]Wykres krzywej ROC dostępny pod adresem https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/



Rysunek 2.17: Pole pod krzywą ROC - AUC - źródło[9]

W zależności od wartości klasyfikatora AUC przyjmuje się następującą interpretację wyniku:

- Gdy AUC = 1, klasyfikator doskonale rozróżnia wszystkie punkty klasy dodatniej i ujemnej
- Gdy AUC = 0 klasyfikator przewiduje wszystkie wyniki negatywne jako pozytywne, a wszystkie pozytywne jako negatywne
- Gdy $0.5 <$ AUC < 1 klasyfikator jest w stanie odróżnić dodatnie wartości klas od ujemnych klas. Dzieje się tak, ponieważ AUC jest w stanie odróżnić więcej liczb i TN od FN i FP
- Gdy AUC = 0.5 klasyfikator nie jest w stanie odróżnić punktów klasy dodatniej od ujemnej. Może to dowodzić o losowości przewidywania klasy, bądź stałą klasę dla wszystkich punktów danych.

Heagerty i Zheng [21] opisują krzywe ROC incydentów dynamicznych (*incident/dynamic* - I/D) zdefiniowanych jako $ROC_t^{I/D}(p)$, gdzie p oznacza dynamiczny współczynnik wskaźników fałszywie pozytywnych, zaś $ROC_t^{I/D}(p)$ charakteryzujemy jako właściwy współczynnik tychże współczynników. Oznaczmy c^p jako próg dający fałszywie pozytywny współczynnik p: $P(M_i > c^p | T_i > t) = 1 - TNR^D(c,t)$.
Używając TP oraz FP funkcji oszacowania

$$TP_t^I(c) = TPR^I(c,t)$$

$$TP_t^D(c) = TNR^D(c,t)$$

krzywą ROC możemy zapisać w postaci złożenia $TP_t^I(c)$ oraz funkcji odwrotnej $[TP_t^D]^{-1}(c) = c^p$ w postaci:

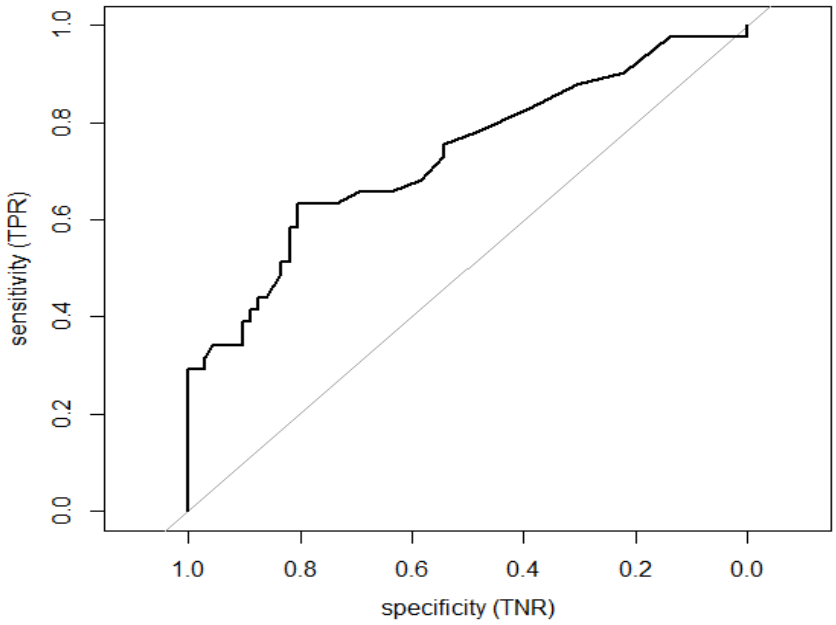
$$ROC_t^{I/D}(p) = TP_t^I(c) \cdot [FP_t^D]^{-1}(p) \tag{2.24}$$

[9]Pole pod krzywą ROC http://sigmaquality.pl/uncategorized/ustawienie-progow-w-modelu-regresji-logistycznej/

Dla $p \in [0, 1]$. Definiujemy także pole pod krzywą ROC zależne od czasu jako:

$$AUC(t) = \int_0^1 ROC_t^{I/D}(p)\, dp \tag{2.25}$$

PRZYKŁAD
W programie `R` za pomocą biblioteki `pROC` wyznaczymy krzywą ROC, a następnie obliczymy pole AUC pod nią. Skorzystamy z danych `aSAH`[10] zawierających informacje na temat 113 pacjentów z krwotokiem podpajęczynówkowym tętniakiem.



Rysunek 2.18: Wykres krzywej ROC - opracowanie własne

Widzimy na powyższym wykresie, że nasza krzywa ROC usytuowana jest ponad szarą linią oznaczającą losową klasyfikację, co może wskazywać na to, że pole AUC będzie lepiej dopasowane, czyli jego wartość będzie bliższa 1.

```
Call:
roc.formula(formula = outcome ~ s100b, data = aSAH)

Data: s100b in 72 controls (outcome Good) < 41 cases (outcome Poor).
Area under the curve: 0.7314
```

Rysunek 2.19: Pole pod krzywą ROC - AUC - opracowanie własne

Wartość AUC = 0.7314, co dowodzi, że nasz klasyfikator jest w stanie odróżnić dodatnie wartości klas od tych ujemnych.

[10]Dane dostępne pod adresem https://rdrr.io/cran/pROC/man/aSAH.html

2.5.1 Time - dependent ROC

Analiza krzywej ROC jest dobrze opracowana do oceny, jak dokładnie marker jest w stanie odróżnić osoby, u których wystąpiła choroba, od osób, u których nie wystąpiła. W klasycznym podejściu stan zdarzenia i wartość markera dla danej osoby są stałe w czasie. Jednakże w praktyce, zarówno stan choroby jak i wartość markera ulegają zmianom. Choroba, która nie wystąpiła wcześniej u danej grupy osób, może rozwinąć się w czasie dłuższej obserwacji, a także wartość ich markera może zmienić się w porównaniu z wartością wejściową podczas obserwacji. Dlatego, do analizy przeżycia bardziej odpowiednia jest krzywa ROC zależna od czasu - *Time - dependent ROC*.

Ustalmy, że T_i jest czasem wystąpienia choroby, zaś X_i jest wejściową wartością markera dla osobnika $i = (1, \ldots, n)$. Definiujemy obserwowany czas zdarzenia $Z_i = min(T_i, C_i)$, gdzie C_i jest czasem ocenzurowania zaś δ_i wskaźnikiem cenzurowania przyjmującym wartość 1 jeśli wystąpi interesujące nas zdarzenie, bądź 0 w przeciwnym razie. Ustalmy, że $D_i(t)$ jest stanem choroby w czasie t przyjmującym wartości 0 lub 1. Dla danego progu c funkcję czułości (2.17) i swoistości (2.19) możemy zapisać jako:

$$TPR = sensitivity(c,t) = P(X > c|D(t) = 1) = \frac{(1 - S(t|X > c))P(X > c)}{1 - S(t)} \tag{2.26}$$

$$TNR = specificity(c,t) = P(X \leq c|D(t) = 0) = \frac{S(t|X \leq c)P(X \leq c)}{S(t)} \tag{2.27}$$

Korzystając z powyższych wzorów krzywą $ROC(t)$ oznaczamy jaką funkcję wykreślającą $TPR(c,t)$ w funkcji $1 - TNR(c,t)$ dla progów c. Zależne od czasu AUC zdefiniowane jest przez:

$$AUC(t) = \int_{-\infty}^{\infty} TPR(c,t)\frac{\sigma(1 - TNR(c,t))}{\sigma c}\, dc \tag{2.28}$$

AUC jest równe prawdopodobieństwu prawidłowego uporządkowania wyników testu diagnostycznego losowo wybranej pary osób zdrowych i chorych.

Heagerty i Zheng [21] zaproponowali trzy różne definicje do szacowana zależnej od czasu czułości i swoistości dla cenzurowanych czasów zdarzeń. Poniżej omówimy je wszystkie.

1. *Cumulative/ dynamic - cumulative sensitivity and dynamic specificity* (C/D)
   W każdym punkcie czasowym t osobnik klasyfikowany jest jako przypadek (osoba, która doświadczyła zdarzenia między punktem początkowym $t = 0$, z czasem t, bądź kontrola (osobnik pozostający wolny od zdarzenia w czasie t.
   Skumulowana wrażliwość (*cumulative sensitivity*) jest prawdopodobieństwem, że marker danej osoby ma wartość większą niż c wśród osób, które doświadczyły zdarzenia przed czasem t.
   Swoistość dynamiczna (*dynamic specificity*) określa prawdopodobieństwo, że marker danej osoby ma wartość mniejszą lub równą c wśród osobników, które nie otrzymały zdarzenia po czasie t.
   Wrażliwość oraz swoistość definiujemy jako:

$$sensitivity^C(c,t) = P(X_i > c|T_i \leq t)$$
$$specificity^D(c,t) = P(X_i \leq c|T_i > t)$$

2. *Incident/ static - incident sensitivity and static specificity* (I/S)
   Przypadek (I/S) określamy jako osobę ze zdarzeniem w czasie t, podczas gdy kontrola jest osobą wolną od wystąpienia zdarzenia przez ustalony okres obserwacji $(0, t^*)$. Punkt końcowy t^* jest z góry ustalony i uważany za wystarczająco długi do zaobserwowania wystąpienia interesującego nas zdarzenia.
   Wrażliwość na incydent (*incident sensitivity*) jest prawdopodobieństwem pozytywnego wyniku testu z markerem w jednostkach czasu t przed zdarzeniem dla osoby mającej czas w T_i.
   Swoistość statyczna (*static specificity*) to prawdopodobieństwo, że dana osoba pozostanie wolna od zdarzeń przez t^* jednostki czasu po zmierzeniu markera.
   Wrażliwość oraz swoistość definiujemy jako:

$$sensitivity^I(c,t) = P(X_i > c|T_i = t)$$

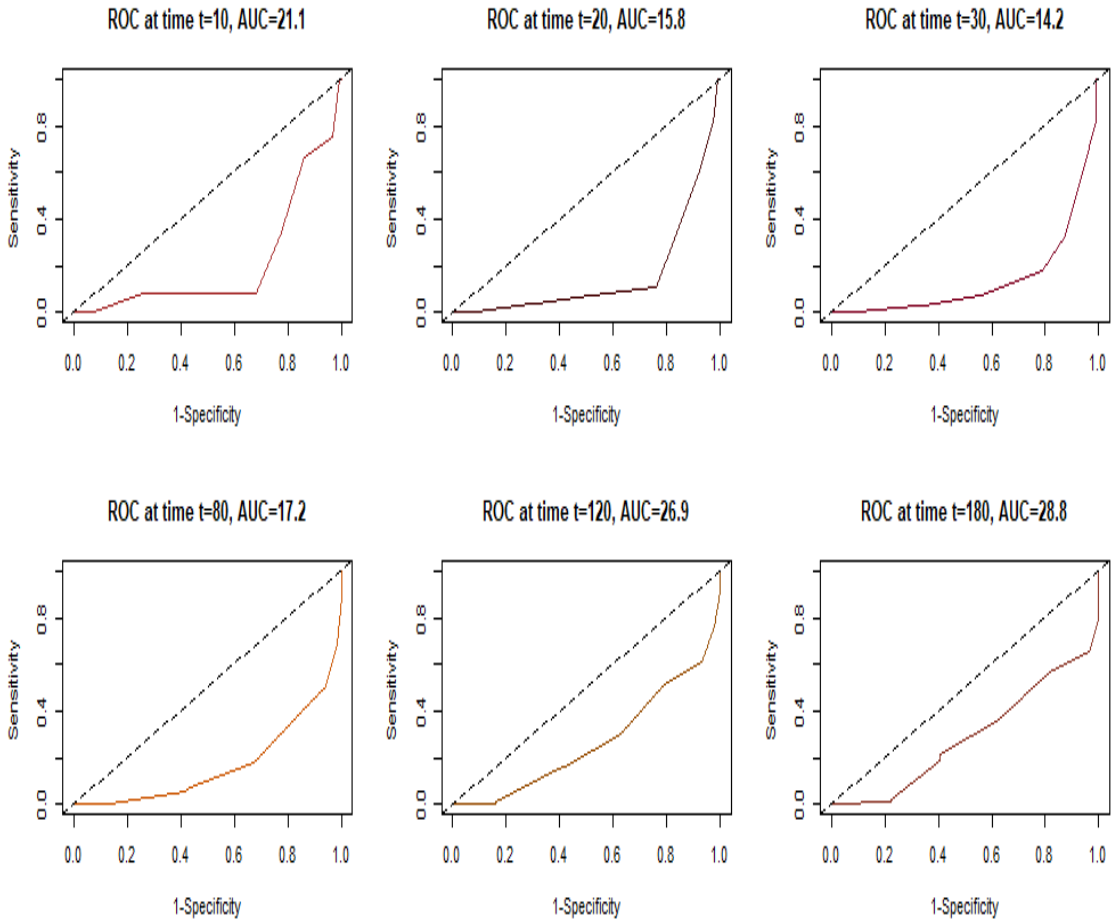
$$specificity^S(c,t^*) = P(X_i \leq c|T_i > t^*)$$

3. *Incident/ dynamic - incident sensitivity and dynamic specificity* (I/D)
   I/D definiujemy jako jednostkę ze zdarzeniem w czasie t, podczas gdy kontrola jest jednostką wolną od zdarzeń w czasie t. Nie występują osobniki kontroli czy przypadku.
   Wrażliwość na incydent (*incident sensitivity*) to prawdopodobieństwo, że dana osoba ma wartość markera większą niż c wśród osób, które doświadczyły zdarzenia w czasie t.
   Swoistość dynamiczna (*dynamic specificity*) to prawdopodobieństwo, że dana osoba ma wartość markera mniejszą niż lub równą c wśród osobników, które nie otrzymały zdarzenia w czasie t
   Wrażliwość oraz swoistość definiujemy jako:

$$sensitivity^I(c,t) = P(X_i > c|T_i = t)$$

$$specificity^D(c,t) = P(X_i \leq c|T_i > t)$$

PRZYKŁAD
Dane `veteran`[11] zawierają informacje na temat leczenia raka płuc dwoma schematami - standardowym oraz testowym. Korzystając z funkcji `timeROC` narysujemy 6 różnych wykresów krzywej ROC[12] oraz wyznaczymy wartości pola AUC dla każdego z nich.



Rysunek 2.20: Wykresy różnych krzywych ROC zależnych od czasu - opracowanie własne

A oto tabelaryczne zestawienie powyższych danych:

Time	Cases	Survivors	Censored	AUC (%)
t = 10	12	123	2	21.07 %
t = 20	27	108	2	15.84 %
t = 80	67	67	3	17.15 %
t = 120	88	43	6	26.88 %
t = 180	103	27	7	28.76 %

Tabela 2.3: Krzywa ROC zależna od czasu

Dla wszystkich wartości dla każdej krzywej ROC zależnej od czasu widzimy inną wartość procentową pola AUC. To znaczy, że zmiana czasu t zmienia wartość AUC.

[11] https://cran.r-project.org/web/packages/survival/survival.pdf, strona 177

[12] Funkcja `timeROC` oszacowuje odwrotne prawdopodobieństwo cenzurowania skumulowanej/dynamicznej krzywej ROC. Funkcja ta oblicza obszary pod krzywymi ROC w zależności od czasu.

Rozdział 3

ROZKŁADY PROBABILISTYCZNE CZASÓW ŻYCIA

Przy modelowaniu czasów życia istotną rolę odgrywają te rozkłady, które są określone na nośniku $(0, \infty)$. My jednak skupimy swoją uwagę na tych rozkładach, które powstają z przekształcenia zmiennych losowych z rozkładu normalnego.

Zmienną losową X z rozkładu normalnego oznaczamy poprzez $X \sim \mathrm{N}(\mu, \sigma^2)$, gdzie $\mu \in R$ oznacza parametr lokalizacji, zaś $\sigma > 0$ jest parametrem skali. Powyższy zapis zawiera też informację o momentach tej zmiennej losowej: μ jest równe wartości oczekiwanej, zaś σ^2 jest wariancją tego rozkładu.

Gęstość rozkładu normalnego $\mathrm{N}(\mu, \sigma^2)$ dana jest wzorem

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right), \ x \in \mathbb{R}.$$

Dystrybuantę rozkładu normalnego $\mathrm{N}(m, \sigma^2)$ w punkcie x będziemy oznaczać symbolem $\Phi(x, m, \sigma)$. Jeżeli $m = 0, \sigma = 1$ to zastosujemy zapis $\Phi(x)$. Do tego niech $P(X > x) = S(x, m, \sigma) = 1 - \Phi(x, m, \sigma)$.

W naszych rozważaniach będziemy też korzystać z wielowymiarowego rozkładu normalnego. Mówimy, że wektor $(X_1, ..., X_n)$ wielowymiarowy rozkład normalny, jeśli istnieją wektor $\mathbf{m} = (m_1, ..., m_n)$ oraz dodatnio określona macierz Σ wymiaru $n \times n$ takie, że gęstość wektora $(X_1, ..., X_n)$ w punkcie $x = (x_1, ..., x_n)$ wyraża się wzorem

$$f(x) = \frac{1}{\sqrt{2\pi}^n \sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x - \mathbf{m})\Sigma^{-1}(x - \mathbf{m})^T\right). \tag{3.1}$$

Piszemy $X \sim \mathrm{N}_n(\mathbf{m}, \Sigma)$. Dystrybuantę wektora normalnego o średniej $\mathbf{m}$ i macierzy kowariancji Σ będziemy oznaczać symbolem $\Phi_n(x, \mathbf{m}, \Sigma)$. W przypadku wektora $\mathbf{m} = (0, ..., 0)$ i $\Sigma = I$ (macierz jednostkowa) zastosujemy zapis $\Phi_n(x)$. Do tego niech $P(X_1 > x_1, ..., X_n > x_n) = S_n(x_1, ..., x_n, \mathbf{m}, \Sigma)$.

3.1 Rozkład Log - Normalny

Rozkład logarytmicznie normalny (zwany także log - normalnym) jest ciągłym rozkładem statystycznym wartości logarytmicznych z powiązanego rozkładu normalnego.

Jeśli zmienna losowa X ma rozkład logarytmicznie - normalny, to zmienna losowa $Y = ln(X) \sim \mathrm{N}(\mu, \sigma^2)$. Zmienna losowa o rozkładzie log - normalnym przyjmuje tylko dodatnie wartości rzeczywiste. Warto podkreślić, że rozkłady te są dodatnio skośne z długimi prawymi ogonami, ze względu na niskie wartości średnie i duże wariancje zmiennych losowych.

Rozkład logarytmiczno - normalny jest wynikiem prac Galtona [13], który uzyskał wyrażenia dla średniej, mediany, mody, wariancji oraz pewnych kwantyli otrzymanego rozkładu. Przechodząc od n niezależnych zmiennych losowych, skonstruował iloczyn, który za pomocą logarytmu przeszedł z iloczynu do sumy nowych zmiennych losowych.

Rozkłady log - normalne najczęściej znajdują zastosowanie w analizie cen akcji. Potencjalne zwroty akcji możemy przedstawić na wykresie o rozkładzie normalnym, zaś ceny akcji na wykresie o rozkładzie logarytmicznie normalnym. Krzywa rozkładu log - normalnego może zostać wykorzystana do dokładniejszej identyfikacji założonego zwrotu, jakiego akcje mogą oczekiwać w określonym czasie.

Rozkład logarytmicznie normalny jest ściśle powiązany z rozkładem normalnym, zatem jego notacja jest identyczna - $ln\mathrm{N}(\mu, \sigma)$, zaś znaczenie tych parametrów całkowicie odmienne (co przedstawia rysunek (3.1)).

Poniższa tabela ukazuje porównanie najważniejszych charakterystyk rozkładów log-normalnego oraz normalnego.

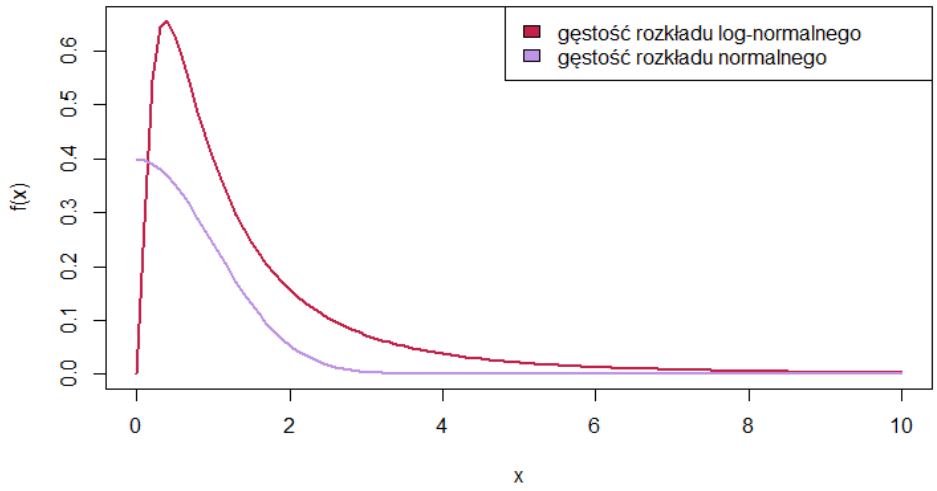
	Rozkład log - normalny	Rozkład normalny
Gęstość	$\frac{1}{\sqrt{2\pi}\sigma x} \cdot \exp\left(\frac{-(\ln x-\mu)^2}{2\sigma^2}\right)$	$\frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$
Dystrybuanta	$\frac{1}{2}\left[1 + erf\left(\frac{\ln(x-\mu)}{\sigma}\right)\right]$	$\frac{1}{2}\left[1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$
Średnia	$\exp\left(\mu + \frac{1}{2}\sigma^2\right)$	μ
Wariancja	$\left(e^{\sigma^2} - 1\right) \cdot e^{2\mu+\sigma^2}$	σ^2

Tabela 3.1: Porównanie charakterystyk rozkładów log - normalnego oraz normalnego.

Gdzie erf oznacza funkcję błędu[1].

Wykres, który przedstawiamy powyżej jest wynikiem porównania gęstości rozkładu normalnego $\mathrm{N}(0, 1)$ oraz rozkładu logarytmicznie normalnego $ln\mathrm{N}(0, 1)$, dla zadanych parametrów $\mu = 0$ oraz $\sigma^2 = 1$. Potwierdza on wcześniejsze wnioski, że nie możemy traktować identycznie parametrów pochodzących z rozkładu normalnego z tymi pochodzącymi z rozkładu logarytmicznie - normalnego.

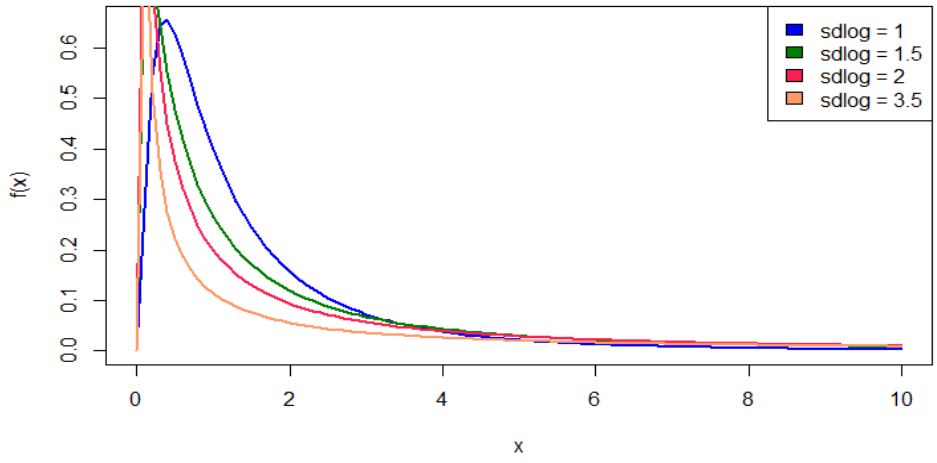
[1]Funkcja błędu występująca podczas całkowania rozkładu normalnego (będącego znormalizowaną formą funkcji Gaussa) jest zdefiniowana jako $erf(z) = \int_0^z \frac{2}{\sqrt{\pi}} e^{-t^2}\, dt$. Warto podkreślić, że niektórzy autorzy (na przykład Whittaker i Watson [43] definiują w swoich artykułach $erf(z)$ bez uwzględnienia przodującego czynnika $\frac{2}{\sqrt{\pi}}$.



Rysunek 3.1: Porównanie wykresów funkcji gęstości rozkładu log-normalnego oraz normalnego - opracowanie własne

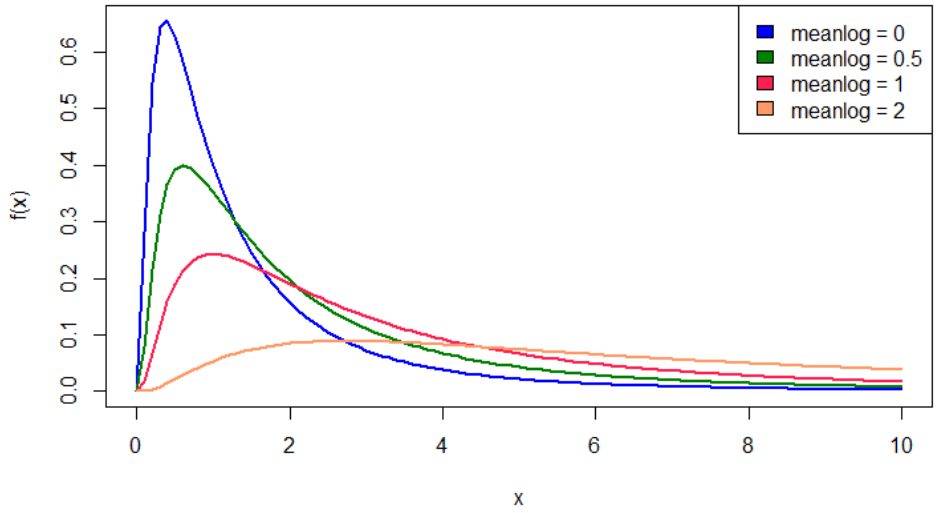
W rozkładzie normalnym 68% wyników mieści się w obrębie jednego odchylenia standardowego, zaś 95% w obrębie dwóch odchyleń standardowych. W centrum mediana oraz średnia są takie same. Rozkład log - normalny różni się od rozkładu normalnego. Jak widać, główna ich różnica dotyczy kształtu. Rozkład normalny jest symetryczny, podczas gdy rozkład log - normalny przyjmuje jedynie wartości dodatnie, przez co nie jest on symetryczny i tworzy krzywą skośną w prawo. Kolejną różnicą wartą uwagi jest fakt, że wartości użyte do uzyskania rozkładu log - normalnego mają rozkład normalny.

Kolejny wykres obrazuje zmienność rozkładu log - normalnego, ze stałym parametrem lokalizacji ($meanlog = 0$) oraz zmiennym parametrem skali ($sdlog$).



Rysunek 3.2: Porównanie wykresów funkcji gęstości rozkładu log-normalnego dla różnych parametrów σ^2 - opracowanie własne

Przyglądając się powyższemu rysunkowi, zauważamy, że wraz ze wzrostem parametru skali, wykres robi się coraz bardziej ostry, stromy.
Zaprezentujemy jeszcze zmienność rozkładu log - normalnego ze stałym parametrem skali ($sdlog = 1$) oraz zmiennym parametrem lokalizacji ($meanlog$)



Rysunek 3.3: Porównanie wykresów funkcji gęstości rozkładu log-normalnego dla różnych parametrów μ - opracowanie własne

Dokładnie widzimy, że wraz ze wzrostem parametru lokalizacji, wykres funkcji gęstości coraz bardziej się spłaszcza.

We wcześniejszych rozdziałach wspominaliśmy już, że rozkład analizy przeżycia podzielony jest na trzy główne funkcje:

- funkcję przeżycia
- funkcję gęstości prawdopodobieństwa
- funkcję hazardu

Jednym z interesujących rozkładów jest rozkład logarytmiczno - normalny służący do modelowania utrzymania systemu. Rozkład ten ma trzy wzorce współczynnika ryzyka: rosnący, malejący, bądź *upside-down bathtub*.

Funkcja gęstości rozkładu log - normalnego jest postaci

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma t} \exp\left[-\frac{1}{2}\left(\frac{\ln(t) - \mu}{\sigma}\right)^2\right], \ t > 0. \tag{3.2}$$

Wykonując całkowanie, dostajemy dystrybuantę:

$$F(t) = \Phi\left[\frac{\ln(t) - \mu}{\sigma}\right]. \tag{3.3}$$

Funkcja przeżycia opisana jest jako:

$$S(t) = 1 - F(t) = 1 - \Phi\left[\frac{\ln(t) - \mu}{\sigma}\right] \tag{3.4}$$

Korzystając z zależności między funkcją gęstości a funkcją przeżycia, otrzymujemy funkcję hazardu rozkładu logistycznie normalnego zadaną wzorem:

$$h(t) = \frac{f(t)}{S(t)} = \frac{\frac{1}{\sqrt{2\pi}\sigma t}\exp\left[-\frac{1}{2}\left(\frac{\ln(t)-\mu}{\sigma}\right)^2\right]}{1 - \Phi\left[\frac{\ln(t)-\mu}{\sigma}\right]} \tag{3.5}$$

3.2 Rozkład Birnbauma – Saundersa

Rozkład Birnbauma – Saundersa jest dwuparametryczną rodziną długości życia prezentowaną poprzez dane dotyczące zmęczenia. Geneza tego rozkładu była motywowana problemami drgań w samolotach komercyjnych, przeznaczonych do wykonywania lotów zarobkowych, powodujących zmęczenie materiałów - stąd również pochodzi jego druga nazwa, czyli rozkład trwałości zmęczeniowej.
Rozkład jest szeroko stosowany w badaniach zmęczenia, niezawodności, a także jest naturalnym modelem w wielu sytuacjach, w których kumulacja pewnego czynnika wymusza przekroczenie przez wymierną charakterystykę progu krytycznego.
Oto kilka możliwych przykładów jego zastosowań.

- Modelowanie frekwencji pierścienic zarówno w drzewostanach o charakterze zbliżonym do pierwotnego jak i w plantacjach
- Jakość powietrza wynikająca z akumulacji ilości zanieczyszczeń w powietrzu w danym okresie czasowym
- Nagromadzenie szkodliwych substancji w płucach z powodu zanieczyszczenia powietrza
- Występowanie przewlekłych chorób serca i różnych rodzajów raka w wyniku skumulowanych uszkodzeń spowodowanych kilkoma czynnikami ryzyka powodującymi degradację i prowadzącymi do procesu zmęczenia

Rozkład Birnbauma - Saundersa opiera się na skumulowanych obciążeniach powodujących zmęczenie materiałów. Rozkład Birnbauma i Saundersa [6] został wyprowadzony z modelu pokazującego, że łączny czas, jaki upłynął do momentu, gdy skumulowane uszkodzenie wywołane rozwojem oraz wzrostem dominującego pęknięcia, przekroczy wartość progową i spowoduje, że próbka materiału przestanie działać. Co ciekawe, Desmond[2] [10] wzmocnił uzasadnienie użycia tego rozkładu, równocześnie rozluźniając poczynione wcześniej niektóre założenia.

Birnbaum i Saunders [6] rozważali jedynie znormalizowane materiały próbek, które zostały poddane zmiennym naprężeniom w wyniku okresowego obciążenia. Poprzez obciążenie rozumiemy ciągłą jednomodalną funkcję na przedziale jednostkowym, której wartość

[2]Desmond [10] przebadał dzieci w wieku przedszkolnym w ośrodku badawczym oraz w szkole, aby określić, w jakim stopniu komunikacja rodzinna pośredniczyła w ich rozumieniu telewizji

w dowolnym momencie daje naprężenie wywołane ugięciem próbki.
Ustalmy, że $l_1, l_2, \ldots$ jest sekwencją obciążeń, które mają zostać przyłożone do każdej oscylacji tak, że przy i-tym obciążeniu oscylacyjnym przyłożone zostanie l_i. Zakładamy, że obciążenie jest funkcją cykliczną oraz ciągłą. Konkretyzujemy, że awaria zmęczeniowa wynika ze wzrostu i przedłużenia dominującego pęknięcia. Przy każdej oscylacji pęknięcie wydłuża się o pewną ilość będącą funkcją losową ze względu na zmienność materiału, wielkość przyłożonego naprężenia i pewną liczbę wcześniejszych obciążeń. Stąd otrzymujemy nasze pierwsze założenie, które można wiarygodnie utrzymać na przykład w przypadku cykli ziemia - powietrze - ziemia w badaniach zmęczenia lotniczego.

1. Przyrostowe rozszerzenie pęknięcia X_i po zastosowaniu i-tej oscylacji jest zmienną losową o rozkładzie zależnym od wszystkich obciążeń rzeczywistych rozszerzeń pęknięcia, które go poprzedziły w tym cyklu.

Z założenia 1. wynika, że niezależnie od tego, jaka jest zależność między kolejnymi przypadkowymi wydłużeniami przypadającymi na oscylację w każdym cyklu, losowe całkowite wydłużenia pęknięć na cykl są niezależne. Stąd tworzymy wniosek drugi:

2. Całkowite rozszerzenie pęknięcia Y_j spowodowane j-tym cyklem jest zmienną losową o średniej μ oraz wariancji σ^2 dla wszystkich $j = 1, 2, \ldots$. Ustalamy także notację

$$W_n = \sum_{j=1}^{n} Y_j \tag{3.6}$$

Gdzie dystrybuanta dla $n = 1, 2, \ldots$ jest postaci

$$H_n(w) = P(W_n \leq w) \tag{3.7}$$

Wynika to z tego, że rozkład C liczby takich cykli do zniszczenia, w przypadku gdy zniszczenie jest zdefiniowane jako długość pęknięcia przekraczającą pewną ustaloną długość krytyczną ω wynosi

$$P(C \leq n) = 1 - H_n(\omega) \tag{3.8}$$

Czyli

$$P(C \leq n) = P\left(\sum_{i=1}^{n} \frac{Y_j - \mu}{\sigma\sqrt{n}} > \frac{\omega - n\mu}{\sigma\sqrt{n}}\right) \tag{3.9}$$

Ponieważ Y_j są niezależnymi zmiennymi o jednakowym rozkładzie, używając Centralnego Twierdzenia Granicznego (CTG), równanie (3.9) dla dużych n można przybliżać rozkładem normalnym standardowym. Stąd otrzymujemy założenie trzecie, które za wartość dokładną przyjmuje przybliżoną równość. Takie założenie można uzasadnić jedynie względami fizycznymi, a nie matematycznymi

3.

$$P(C \leq n) = 1 - \Phi\left(\frac{\omega - n\mu}{\sigma\sqrt{n}}\right) = \Phi\left(\frac{\mu\sqrt{n}}{\sigma} - \frac{\omega}{\sigma\sqrt{n}}\right) \tag{3.10}$$

Dla uproszczenia zakładamy dalej, że rozkłady rozszerzenia pęknięcia są identyczne dla wszystkich cykli. Niech

$$\alpha = \frac{\sigma}{\sqrt{\mu\omega}}, \qquad \beta = \frac{\omega}{\mu}. \tag{3.11}$$

Parametr n zamienimy na nieujemną zmienną rzeczywistą $t > 0$. Jeśli teraz oznaczymy ciągłe rozszerzenie zmiennej losowej C poprzez T (ciągłą, nieujemną zmienną losową), to z wniosku 3. oraz z (3.11) otrzymujemy dystrybuantę zmiennej losowej T. Stosować będziemy oznaczenie $T \sim \mathrm{BS}(\alpha, \beta)$. Na mocy (3.10) uzyskujemy postać dystrybuanty zmiennej $T \sim \mathrm{BS}(\alpha, \beta)$. Dana jest ona w postaci:

$$F(t; \alpha, \beta) = \Phi\left(\alpha^{-1}\gamma\left(\frac{t}{\beta}\right)\right), \qquad t, \alpha, \beta > 0, \qquad \gamma(t) = t^{1/2} - t^{-1/2}. \tag{3.12}$$

Powyższa, dwuparametrowa rodzina rozkładów jest modelem rozkładu życia zmęczeniowego. Jest ona stosowana do modelowania czasu życia zmęczeniowego obok takich rozkładów jak Weibulla (znany ze swoich zastosowań jako rozkład długości żywotności zmęczeniowej), czy Gamma (uzyskany jako rozkład trwałości). Dostrzegamy, że T jest dwuparametrową zmienną losową, z parametrem β jako parametrem lokalizacji, bądź skali, oraz α oznaczającym parametr kształtu. Powszechniej spotykaną formą dystrybuanty (3.12) jest:

$$F(t; \alpha, \beta) = \Phi\left[\frac{1}{\alpha}\left(\left(\frac{t}{\beta}\right)^{1/2} - \left(\frac{\beta}{t}\right)^{1/2}\right)\right] \tag{3.13}$$

Rozkład Birnbauma - Saundersa jest unimodalny z medianą β. Cztery pierwsze momenty zwykłe wynoszą odpowiednio:

- $ET = \beta\left(1 + \frac{1}{2}\alpha^2\right)$
- $E(T^2) = \beta^2\left(1 + 2\alpha + 2\alpha^2 + \frac{3}{2}\alpha^4\right)$
- $E(T^3) = \frac{\beta^3}{2}\left(2 + 9\alpha^2 + 18\alpha^4 + 15\alpha^6\right)$
- $E(T^4) = \frac{\beta^4}{2}\left(2 + 16\alpha^2 + 60\alpha^4 + 120\alpha^6 + 105\alpha^8\right)$

Stąd otrzymujemy charakterystyki liczbowe rozkładu BS:

- wariancja: $\sigma^2 = (\alpha\beta)^2\left(1 + \frac{5\alpha^2}{4}\right)$
- współczynnik skośności: $\gamma = \frac{4\alpha(11\alpha^2+6)}{(5\alpha^2+4)^{3/2}}$
- kurtoza: $\kappa = 3 + \frac{6\alpha^2(93\alpha^2+40}{(5\alpha^2+4)^2}$

Obserwujemy, że dla ustalonego parametru α wariancja T wzrasta wraz ze wzrostem parametru skali β.

Fakt 3.1. *Saunders [39] opisał, że jeżeli* $T \sim \mathrm{BS}(\alpha, \beta)$ *to* $T^{-1} \sim \mathrm{BS}(\alpha, \frac{1}{\beta})$. *Ponadto, dla każdej rzeczywistej liczby* $a > 0$, aT *ma rozkład* $\mathrm{BS}(\alpha, a\beta)$.

Powyższy fakt pokazuje, że parametr β w rzeczywistości jest współczynnikiem skali dla dowolnego rozkładu normalnego. W wielu zastosowaniach nazywa się charakterystycznym życiem. Ponadto określamy, że rozkład Birnbauma - Saundersa jest niezmienny pod wzajemną transformacją.

Fakt 3.2. *Niech* $X \sim \mathrm{N}(0, \alpha^2/4))$. *Wtedy zmienną* $T \sim \mathrm{BS}(\alpha, \beta)$ *możemy zapisać jako transformację:*

$$T = \beta(1 + 2X^2 + 2X(1 + X^2)^{1/2}) \tag{3.14}$$

Dowód. Wcześniej określiliśmy, że $\alpha^{-1}\gamma\left(\frac{T}{\beta}\right)$ jest standardową normalną zmienną losową ze średnią zero i wariancją jednostkową. Ustalmy, że $X \sim \mathrm{N}(0, \frac{\alpha^2}{4})$, stąd

$$2X = \gamma\left(\frac{T}{\beta}\right)$$

Dla $x \in \mathbb{R}$ definiujemy nową funkcję

$$\begin{aligned}
&\psi = \gamma^{-1}(2x)\\
&T = \beta\psi(X)\\
&\psi(x) = [x + (x^2 + 1)^{1/2}]
\end{aligned}$$

Dalej, korzystając z faktu, że $T = \beta\psi(X)$, otrzymujemy oczekiwaną tezę:

$$T = \beta(1 + 2X^2 + 2X(1 + X^2)^{1/2})$$

□

Podczas gdy wiadomo, że każda zmienna losowa o rozkładzie określonym przez (3.8), dla której Y_j są nieujemne, mają rosnący wskaźnik awaryjności, nasza zmienna losowa T nie ma tej własności - jego średni wskaźnik awaryjności prawie nie maleje. Typowe dane dotyczące zmęczenia, pokazują, że nie można zakładać, że wskaźnik awaryjności zawsze rośnie.

Fakt 3.3. *Niech* $T \sim \mathrm{BS}(\alpha, \beta)$. *Wtedy zmienna losowa* $N = \frac{1}{\alpha}\left(\sqrt{\frac{T}{\beta}} - \sqrt{\frac{\beta}{T}}\right) \sim \mathrm{N}(0, 1)$

Dowód.

$$\begin{aligned}
&T - \beta - 2\beta X^2 = 2\beta X\sqrt{1 + X^2}\\
&4\beta^2X^4 + 4\beta^2X^2 + \beta^2 - 4\beta TX^2 - 2\beta T + T^2 = 4\beta^2X^2 + 4\beta^2X^4\\
&\frac{1}{4\beta T}(T - \beta)^2 = X^2
\end{aligned}$$

Stąd rozkład X możemy zapisać jako transformację

$$X = \sqrt{\frac{1}{\beta T}} \cdot \frac{T - \beta}{2} \tag{3.15}$$

Co po uwzględnieniu własności rozkładu normalnego prowadzi do tezy. □

Fakt 3.4. *Niech* $T \sim \mathrm{BS}(\alpha, \beta)$. *Wtedy funkcja gęstości prawdopodobieństwa* T, *dla* $t, \alpha, \beta > 0$, *zadana jest wzorem:*

$$f(\alpha, \beta, t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2\alpha^2}\left(\frac{t}{\beta} + \frac{\beta}{t} - 2\right)\right) \frac{1}{2\alpha\beta}\left(\left(\frac{t}{\beta}\right)^{-1/2} + \left(\frac{t}{\beta}\right)^{-3/2}\right) \tag{3.16}$$

$$f(\alpha, \beta, t) = \frac{1}{\sqrt{8\pi}} \exp\left(\frac{1}{\alpha^2}\right) \exp\left(-\frac{1}{2\alpha^2}\left(\frac{t}{\beta} + \frac{\beta}{t}\right)\right) \frac{t^{-3/2}}{\alpha\beta^{1/2}}(t + \beta) \tag{3.17}$$

Dwa przydatne wskaźniki w analizie przeżycia, to funkcja przeżycia (1.1) oraz funkcja hazardu (1.2). Korzystając z równań (3.13) oraz (3.16) otrzymujemy odpowiednio funkcję niezawodności oraz wskaźnik awaryjności $T \sim \mathrm{BS}(\alpha, \beta)$:

$$S(t, \alpha, \beta) = 1 - F(t, \alpha, \beta) = \Phi\left(-\frac{1}{\alpha}\left(\sqrt{\frac{t}{\beta}} - \sqrt{\frac{\beta}{t}}\right)\right) \tag{3.18}$$

$$h(t, \alpha, \beta) = \frac{f(t, \alpha, \beta)}{S(t, \alpha, \beta)} = \frac{\phi\left(\frac{1}{\alpha}\gamma\left(\frac{t}{\beta}\right)\right) \frac{1}{\alpha}\gamma'\left(\frac{t}{\beta}\right)}{\Phi\left(-\frac{1}{\alpha}\gamma\left(\frac{t}{\beta}\right)\right)} \tag{3.19}$$

Gdzie $\gamma(t) = t^{1/2} - t^{-1/2}$ dla $u > 0$.

Analizując powyższe, możemy wyciągnąć trzy wnioski dotyczące funkcji hazardu rozkładu Birnbauma - Saundersa:

- $h(t, \alpha, \beta)$ jest unimodalny dla dowolnej wartości α, rosnący dla $t < t_c$ oraz malejący dla $t > t_c$, gdzie t_c oznacza punkt zmiany $h(t, \alpha, \beta)$
- $h(t, \alpha, \beta)$ zbliża się do $\frac{1}{2\alpha^2\beta}$ dla $t \to \infty$
- $h(t, \alpha, \beta)$ ma tendencje do wzrostu, jak $\alpha \to 0$

Wprowadzenie rodziny rozkładów oparte na wiarygodnych rozważaniach fizycznych nie jest samo w sobie rozstrzygającym argumentem, że taka konkretna rodzina powinna zostać uwzględniona w badaniach życia. Zawsze najpierw musimy skonfrontować badania trwałości zmęczeniowej z rzeczywistymi danymi zmęczeniowymi uzyskanymi w różnych warunkach.

Rozkład Birnbauma–Saundersa jest zaimplementowany w języku programowania R w bibliotece o nazwie `extraDistr`[3]. Zapewnia ona między innymi gęstość, funkcję rozkładu, generator liczb losowych czy funkcję kwantylową. Model BS oparty jest tam na rozkładzie normalnym o gęstości z parametrem α (parametr kształtu) i β zadanej wzorem:

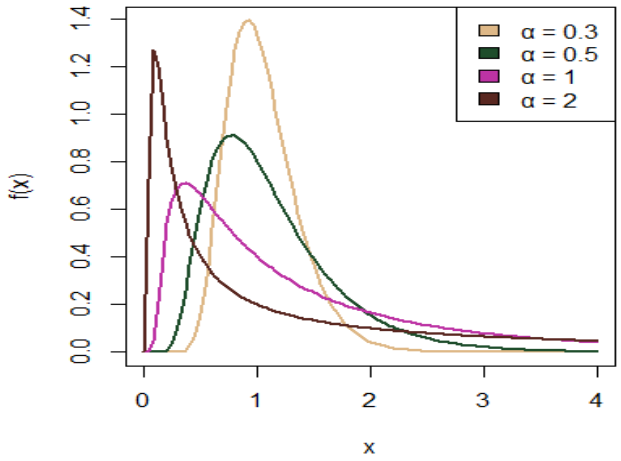
$$f(x) = \left(\frac{\sqrt{\frac{x-\mu}{\beta}} + \sqrt{\frac{\beta}{x-\mu}}}{2\alpha(x-\mu)}\right) \phi\left(\frac{1}{\alpha}\left(\sqrt{\frac{x-\mu}{\beta}} - \sqrt{\frac{\beta}{x-\mu}}\right)\right)$$

[3]https://cran.r-project.org/web/packages/extraDistr/extraDistr.pdf, strony od 10 do 11.

Dystrybuanta opisana jest tam jako:

$$F(x) = \Phi\left(\frac{1}{\alpha}\left(\sqrt{\frac{x-\mu}{\beta}} - \sqrt{\frac{\beta}{x-\mu}}\right)\right)$$

Gdzie $\phi(x)$ oznacza gęstość rozkładu standardowego normalnego.



Rysunek 3.4: Wykres funkcji gęstości rozkładu BS dla wskazanej wartości α przy $\beta = 1$ - opracowanie własne

Rysunek powyżej przedstawia wykresy funkcji gęstości prawdopodobieństwa rozkładu Birnbauma - Saundersa dla różnych wartości jego parametru kształtu α, przy uwzględnieniu jednego parametru skali $\beta = 1$. Na podstawie tego rysunku zauważamy, że rozkład Birnbauma - Saundersa jest ciągły, jednomodalny oraz zawsze dodatni.

Rozdział 4

ANALIZA DANYCH W OPARCIU O MODELE TIME-ROC

Heagerty oraz Zheng [21] dokonali przykładowej analizy koncepcji krzywej ROC zależnej od czasu. Rozważyli przypadek, gdzie marker M (rozkładu Gauss'owskiego) oraz logarytm czasu przeżycia $ln(T)$ (czasy pochodzą z rozkładu log - normalnego) postępują zgodnie z dwuwymiarowym rozkładem normalnym. Wyszli oni z założenia, że wyższa wartość markera wskazuje na wcześniejsze wystąpienie choroby. Z tego powodu zbadali oni rozkłady dwuwymiarowe z ujemną korelacją występującą między markerem a ln(czasu).

W tym rozdziale uogólnimy ich wynik. Przypuśćmy, że wektor $[M, \xi(T)]$ ma dwuwymiarowy rozkład normalny ze średnią $(0,0)$ oraz macierzą kowariancji $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, gdzie ρ jest współczynnikiem korelacji między zmiennymi brzegowymi. O funkcji $\xi : (0,\infty) \to \mathbb{R}$ zakładamy, że jest rosnąca, ciągła i różniczkowalna. Wtedy też istnieje ξ^{-1}.

Gęstość wektora $[M, \xi(T)]$ możemy zapisać jako

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}\frac{1}{1-\rho^2}(x^2 - 2\rho xy + y^2)\right). \tag{4.1}$$

Korzystając z przekształcenia wektorów losowych znajdujemy postać gęstości wektora $[M, T]$. Niech $c(x,y) = x, t(x,y) = \xi^{-1}(y)$. Wtedy szukana gęstość ma postać:

$$g(c,t) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}\frac{1}{1-\rho^2}(c^2 - 2\rho c\xi(t) + \xi^2(t))\right)\xi'(t). \tag{4.2}$$

Zależna od czasu czułość incydentu (*incident sensitivity*) oraz skumulowana swoistość (*cumulative specificity*) zadane są odpowiednio wzorami:

$$P(M > c|T = t) = TP_t^{\mathbb{I}}(c)$$
$$P(M \leq c|T > t) = FP_t^{\mathbb{D}}(c).$$

Znajdziemy gęstość brzegową zmiennej M pod warunkiem $T = t$.

$$g_{M|T=t}(c) = \frac{g(c,t)}{\int_{\mathbb{R}} g(c,t)dc} = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}\frac{1}{1-\rho^2}(c - \rho\xi(t))^2\right). \tag{4.3}$$

Stąd zmienna $M|T = t$ ma rozkład $\mathrm{N}(\rho\xi(t), 1-\rho^2)$. Zatem otrzymujemy

$$P(M > c|T = t) = TP_t^{\mathbb{I}}(c) = \Phi\left(\frac{\rho\xi(t) - c}{\sqrt{1-\rho^2}}\right). \tag{4.4}$$

Analogicznie otrzymujemy postać drugiej interesującej nas funkcji

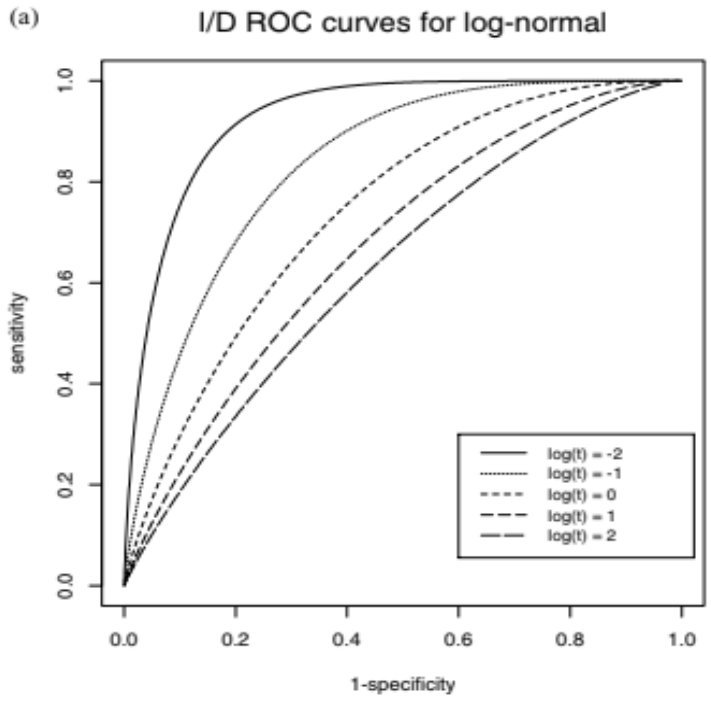
$$P(M \leq c|T > t) = FP_t^{\mathbb{D}}(c) = \frac{S_2(c, \xi(t), (0,0), \Sigma)}{\Phi(-\xi(t))}, \tag{4.5}$$

4.1 Analiza danych z rozkładu Log - normalego

Niech $\xi(t) = \ln(t)$. Przy takich samych założeniach jak powyżej $[M, \ln(T)]$ ma dwuwymiarowy rozkład normalny, ze średnią $(0,0)$ oraz macierzą kowariancji $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, gdzie ρ jest współczynnikiem korelacji między zmiennymi brzegowymi. Zależna od czasu czułość incydentu (*incident sensitivity*) oraz skumulowana swoistość (*cumulative specificity*) zadane są odpowiednio wzorami:

$$P(M > c|T = t) = TP_t^{\mathbb{I}}(c) = \Phi\left(\frac{\rho\ln(t) - c}{\sqrt{1-\rho^2}}\right)$$
$$P(M \leq c|T > t) = FP_t^{\mathbb{D}}(c) = \frac{S_2(c, \ln(t), (0,0), \Sigma)}{\Phi(-\ln(t))}$$

Wykres poniżej przedstawia krzywe I/D ROC dla $\rho = -0.8$ oraz różnych wartości parametru t.



Rysunek 4.1: Wykres krzywych I/D ROC - źródło [21]

Przedstawimy teraz analizę krzywych ROC, gdzie zmienne brzegowe są niezależne, to znaczy, że $\rho = 0$. Ustalamy także, że $\xi(t) = ln(t)$, zaś macierz kowariancji jest postaci $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Stąd, korzystając ze wzorów (4.4) oraz (4.5) definiujemy funkcje czułości oraz swoistości następująco:

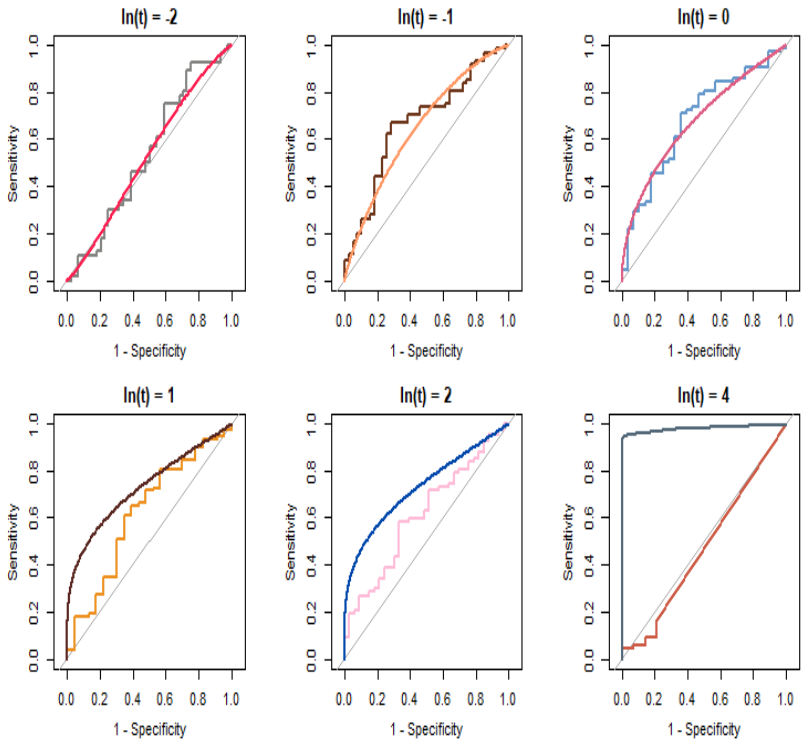
$$TP_t^{\mathbb{I}}(c) = \Phi\left(\frac{-c}{1}\right)$$
$$FP_t^{\mathbb{D}}(c) = \frac{S_2(c, ln(t), (0,0), \Sigma)}{\Phi(-ln(t))}$$

Wylosujemy teraz próbkę 100 wartości dla czasów pochodzących z rozkładu logarytmicznie normalnego oraz markera pochodzącego z czasu z rozkładu normalnego.

```
n_samples <- 100
time <- sort(rlnorm(n = n_samples, mean = 0, sd = exp(-2)))
status <- ifelse(test = (rnorm(n = n_samples) <
                           (rank(time)/n_samples)),
                 yes = 1, no = 0)
```

Rysunek 4.2: Funkcja losująca parametry *time* oraz *status* - opracowanie własne

Następnie wybranych zostało 6 różnych czasów z rozkładu log - normalnego dla stałej wartości średniej ($mean = 0$) w celu zbadania pokrycia się estymowanych wartości krzywych wraz z krzywymi teoretycznymi.

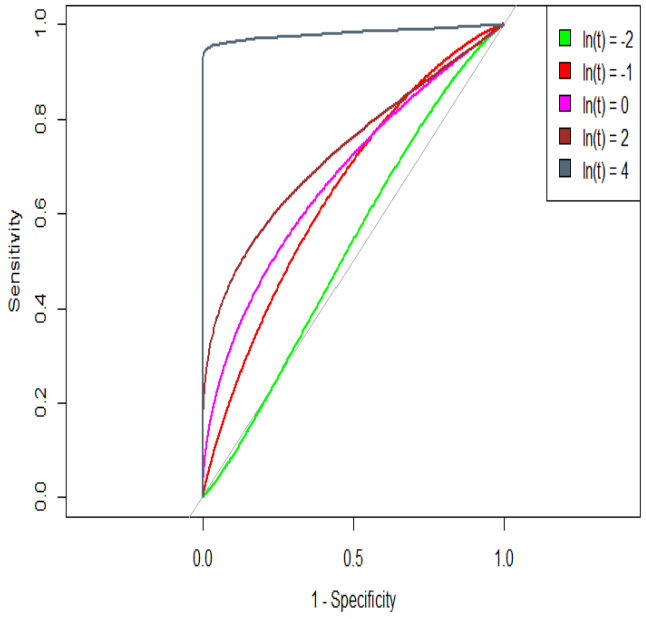


Rysunek 4.3: Estymowana oraz teoretyczna wartość krzywych ROC dla różnych parametrów t, oraz stałej wartości $mean = 0$ - opracowanie własne

AUC(exp(-2))	AUC(exp(-1))	AUC(exp(0))	AUC(exp(2))	AUC(exp(4))
0.5286	0.6505	0.6795	0.7327	0.9818

Tabela 4.1: Porównanie wartości pola AUC(t) dla różnych t - opracowanie własne

Widzimy, że najlepsze dopasowanie krzywych empirycznych i teoretycznych występuje dla $ln(t) = 0$, czyli $t = 1$. Dla coraz większych wartości t krzywe coraz gorzej się pokrywają, bądź jak w przypadku $t = e^4$ - wcale. Wnioskujemy zatem, że w tym przypadku wartości parametru t powinny być jak najbliższe wartości 0.



Rysunek 4.4: Wykres krzywych empirycznych I/D ROC - opracowanie własne

4.2 Analiza danych z rozkładu Birnbauma–Saundersa

Przypomnijmy, że jeśli zmienna losowa T podąża za rozkładem z parametrami kształtu $\alpha > 0$ i parametrem skali $\beta > 0$, to stosujemy notację $T \sim \mathrm{BS}(\alpha, \beta)$. Z faktu 3.3 wiemy, że:

$$N = \frac{1}{\alpha}\left(\sqrt{\frac{T}{\beta}} - \sqrt{\frac{\beta}{T}}\right) \sim \mathrm{N}(0,1)$$

Wynika z tego, że zmienne losowe $T \sim \mathrm{BS}(\alpha, \beta)$ oraz $N \sim \mathrm{N}(0,1)$ są powiązane ze sobą transformacją monotoniczną. W rezultacie stwierdzamy, że dowolną zmienną losową T o rozkładzie Birnbauma - Saundersa można uzyskać jako przekształcenie innej zmiennej losowej N o standardowym rozkładzie normalnym. Podsumowując:

$$T = \beta\left(\frac{\alpha N}{2} + \sqrt{\left(\frac{\alpha N}{2}\right)^2 + 1}\right)$$

Przejdźmy teraz do analizy danych z rozkładu Birnbauma - Saundersa. Ustalamy następującą notację:

- M określa marker diagnostyczny. Umownie ustalamy, że wyższa wartość markera bardziej wskazuje na chorobę
- T oznacza czas niepowodzenia
- C oznacza czas cenzurowania
- $X = min(T, C)$ określa czas obserwacji
- Δ definiujemy jako wskaźnik cenzurowania. Dla $T \leq C, \Delta = 1$, analogicznie dla $T > C, \Delta = 0$

W omawiamym przypadku marker będzie pochodził z rozkładu normalnego, zaś czasy z rozkładu Birnbauma - Saundersa. Ustalmy, że $\xi(t) = \frac{1}{\alpha}\gamma(t)$, gdzie $\gamma(t) = t^{1/2} - t^{-1/2}$, oraz $t = \frac{T}{\beta}$, gdzie $T \sim \text{BS}(\alpha, \beta)$. Wtedy wektor $\left[M, \frac{1}{\alpha}\gamma(t)\right]$ ma dwuwymiarowy rozkład normalny ze średnią $(0, 0)$ oraz macierzą kowariancji $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, gdzie ρ określa współczynnik korelacji między zmiennymi brzegowymi. Gęstość takiego rozkładu określona jest wzorem:

$$g(c,t) = \frac{1}{2\pi\sqrt{1-\rho^2}} exp\left(-\frac{1}{2}\frac{1}{1-\rho^2}\left(c^2 - 2\rho c\frac{1}{\alpha}\gamma(t) + \frac{1}{\alpha^2}\gamma^2(t)\right)\right)\frac{1}{2\alpha}\left(t^{-1/2} + t^{-3/2}\right)$$

Czułość incydentu oraz skumulowaną swoistość, na podstawie wzorów (4.4) oraz (4.5), definiujemy kolejno:

$$TP_t^{\mathbb{I}}(c) = \Phi\left(\frac{\rho\frac{1}{\alpha}\gamma(t) - c}{\sqrt{1-\rho^2}}\right)$$
$$FP_t^{\mathbb{D}}(c) = \frac{S_2(c, \frac{1}{\alpha}\gamma(t), (0,0), \Sigma)}{\Phi\left(-\frac{1}{\alpha}\gamma(t)\right)}$$

Na podstawie wzoru (3.14) stworzyliśmy własną funkcję do generowania zmiennych losowych z rozkładu Birnbauma - Saundersa.

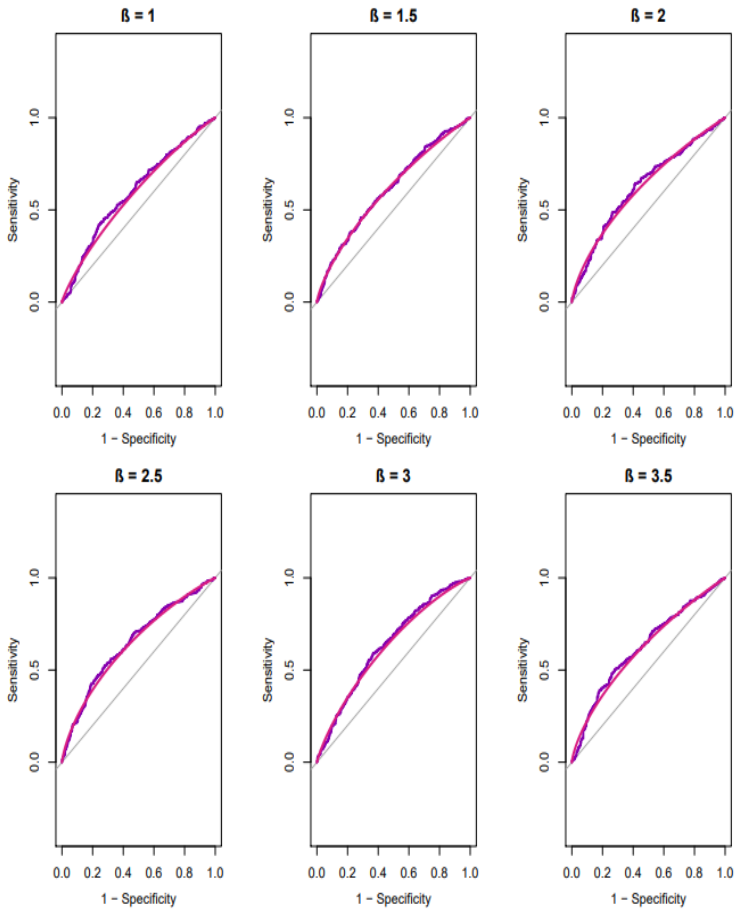
```
rozklad_BS <- function(mu, a, b, n){
  X <- rnorm(n, mu, a^2/4)
  T <- b * (1 + 2*X^2 + 2*X*(1 + X^2)^(1/2))
  return(T)
}
```

Rysunek 4.5: Funkcja generująca zmienne z rozkładu BS - opracowanie własne

Losujemy teraz 1000 liczb z rozkładu Birnbauma - Saundersa, które oznaczamy jako *time*, a następnie tyle samo liczb z rozkładu normalnego, które zapiszemy jako zmienną *status*. Najpierw skupimy się na rozkładzie o stałym parametrze $\alpha = 0.5$ oraz zmienną wartością parametru β.

```
n <- 1000
mu <- 0
a <- 0.5
b <- b
time <- sort(rozklad_BS(mu, a, b, n))
status <- ifelse(test = (rnorm(n = n) < (rank(time)/n)), yes = 1, no = 0)
```

Rysunek 4.6: Funkcja generująca zmienne z rozkładu BS ze stałym parametrem $\alpha = 0.5$ - opracowanie własne

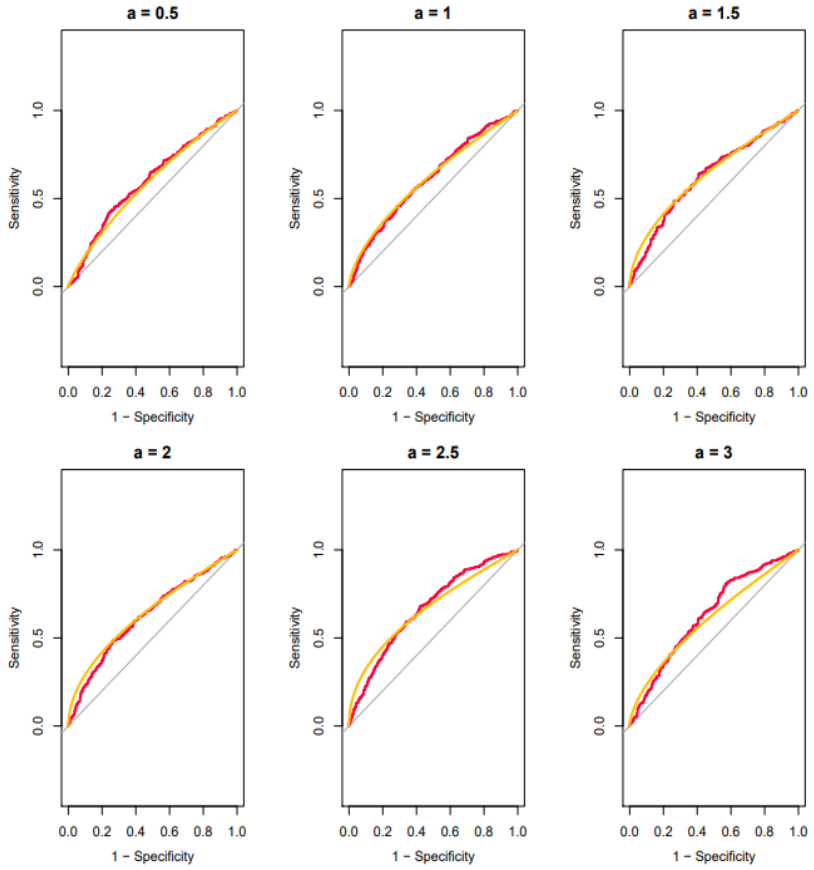


Rysunek 4.7: Estymowana oraz teoretyczna wartość krzywych ROC dla różnych parametrów β, oraz stałej wartości parametru $\alpha = 0.5$ - opracowanie własne

Jak widać na powyższym rysunku, zmienność parametru β nie wpływa na dopasowanie krzywych ROC empirycznych oraz teoretycznych. Zbadamy teraz, czy dla zmiennego parametru α oraz stałego $\beta = 1$ otrzymamy taki sam wniosek.

```
n <- 1000
mu <- 0
a <- a
b <- 1
time <- sort(rozklad_BS(mu, a, b, n))
status <- ifelse(test = (rnorm(n = n) < (rank(time)/n)), yes = 1, no = 0)
```

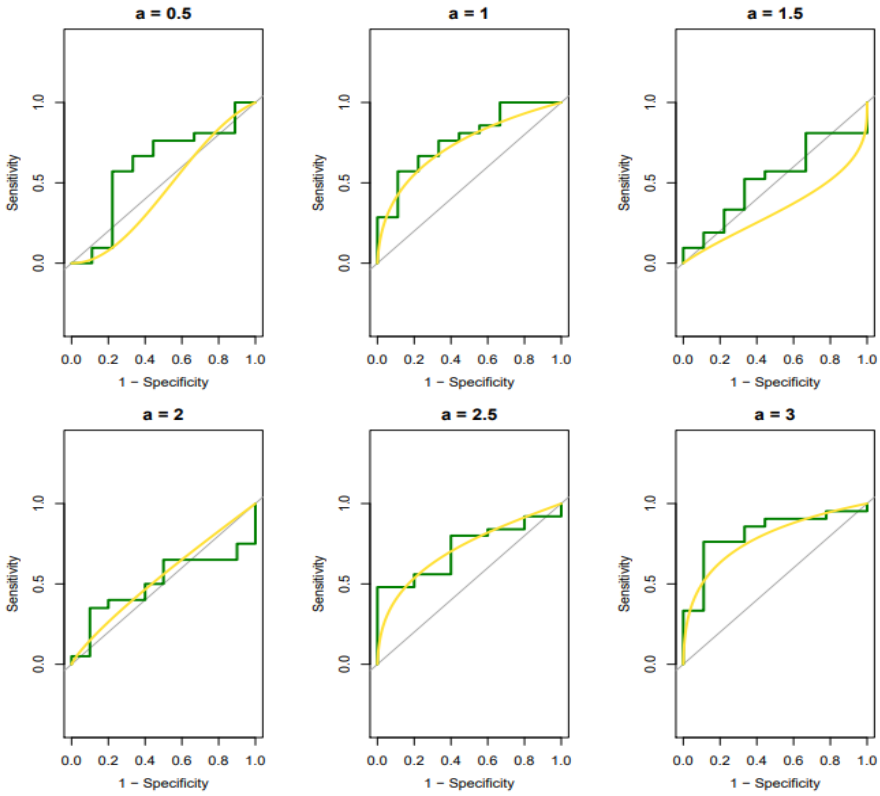
Rysunek 4.8: Funkcja generująca zmienne z rozkładu BS ze stałym parametrem $\beta = 1$ - opracowanie własne



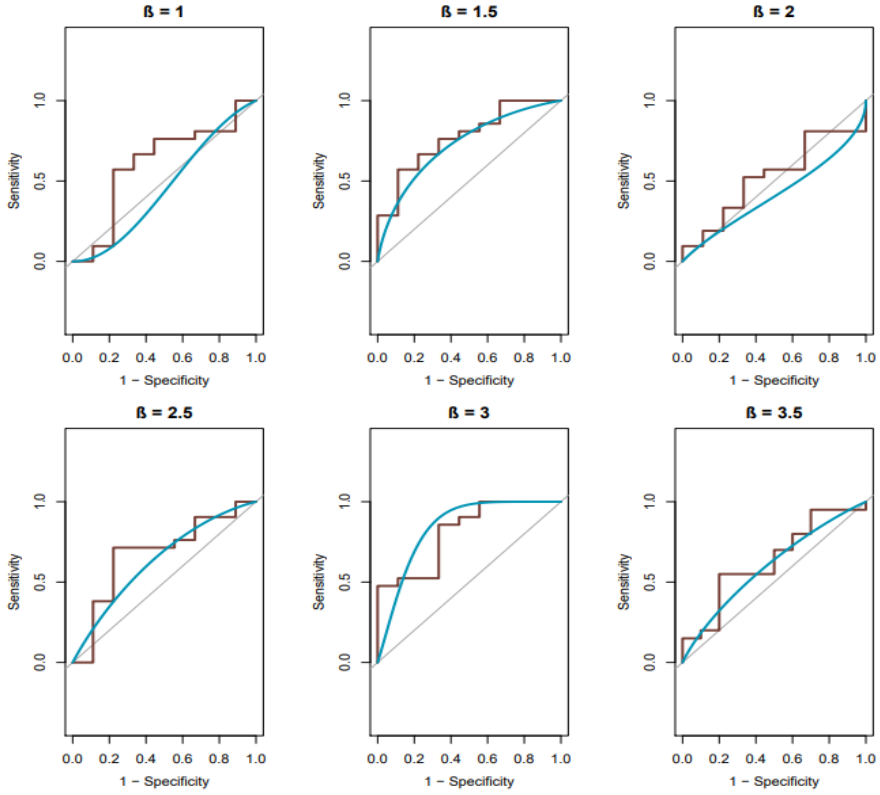
Rysunek 4.9: Estymowana oraz teoretyczna wartość krzywych ROC dla różnych parametrów α, oraz stałej wartości parametru $\beta = 1$ - opracowanie własne

Podsumowując powyższe rozważania, można stwierdzić, że na dopasowanie I/D krzywych ROC, dla dużej próby, krzywych empirycznych oraz teoretycznych nie ma wpływu zmiana parametru kształtu, czy skali.

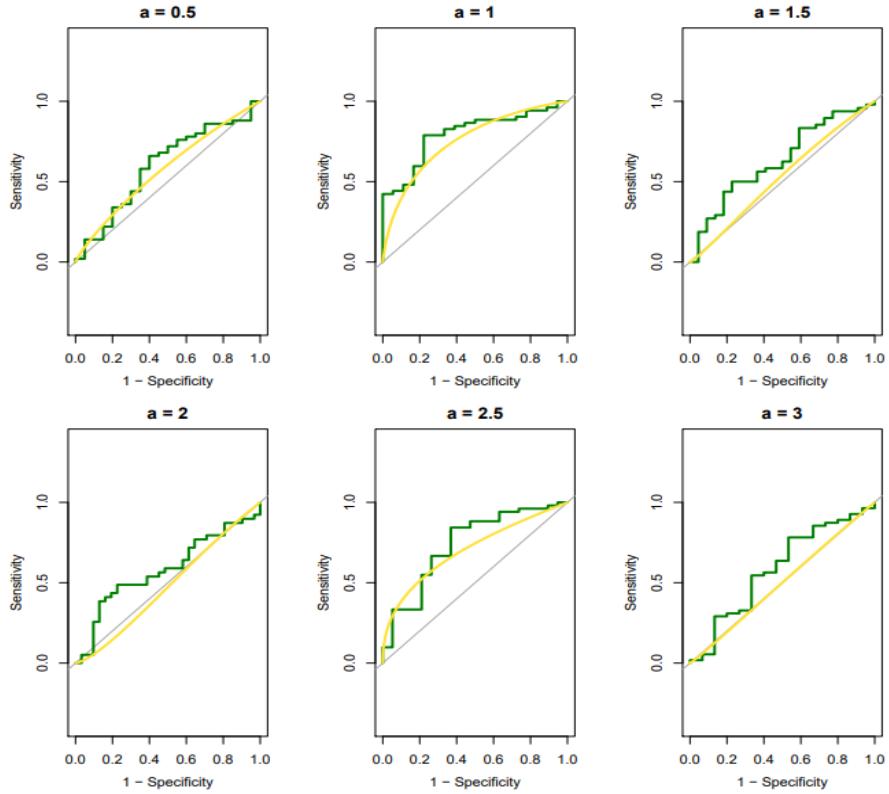
Wiemy już, co dzieje się dla dużych prób losowych. Warto sprawdzić, czy dla małej próby (na przykład $n = 30, 70, 150$), zmiana parametrów α oraz β również nie będzie wpływać na dopasowanie krzywych empirycznych i teoretycznych. Zaczniemy od najmniejszej wybranej próby, czyli $n = 30$, później $n = 70$, na końcu przeanalizujemy wykresy dla $n = 150$.



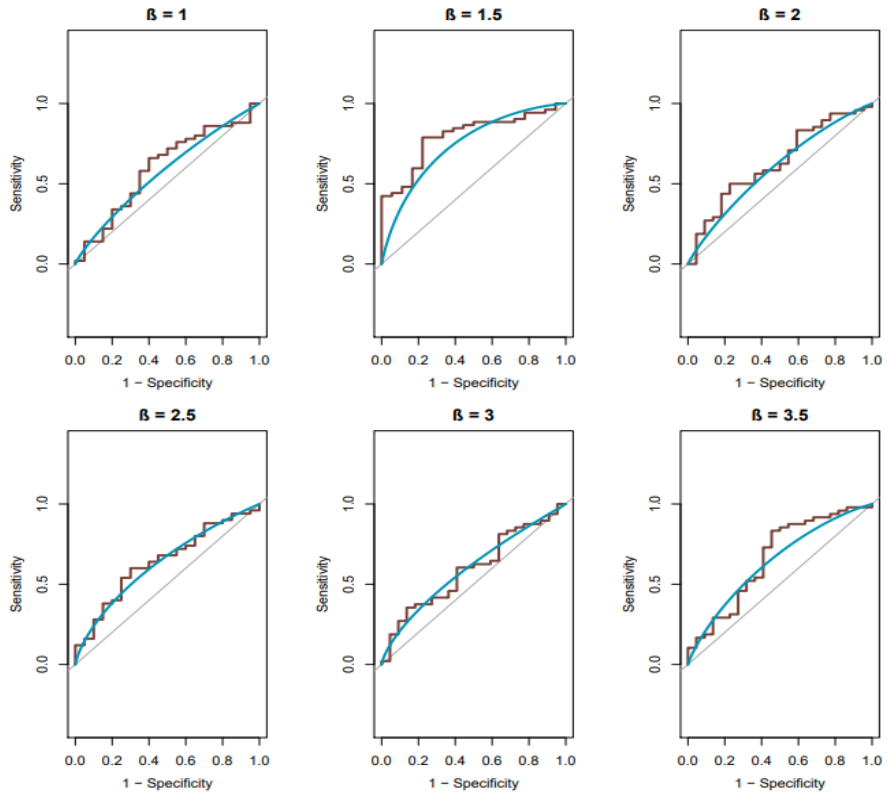
Rysunek 4.10: Estymowana oraz teoretyczna wartość krzywych ROC dla różnych parametrów α, oraz stałej wartości parametru $\beta = 1$ dla próby $n = 30$ - opracowanie własne



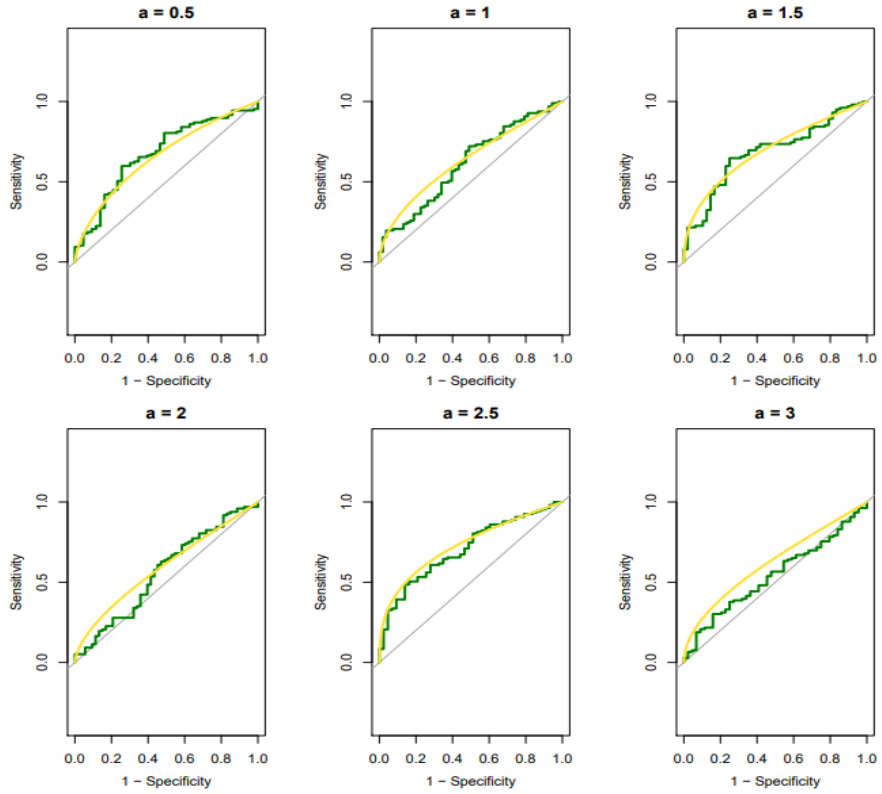
Rysunek 4.11: Estymowana oraz teoretyczna wartość krzywych ROC dla różnych parametrów β, oraz stałej wartości parametru $\alpha = 0.5$ dla próby $n = 30$ - opracowanie własne



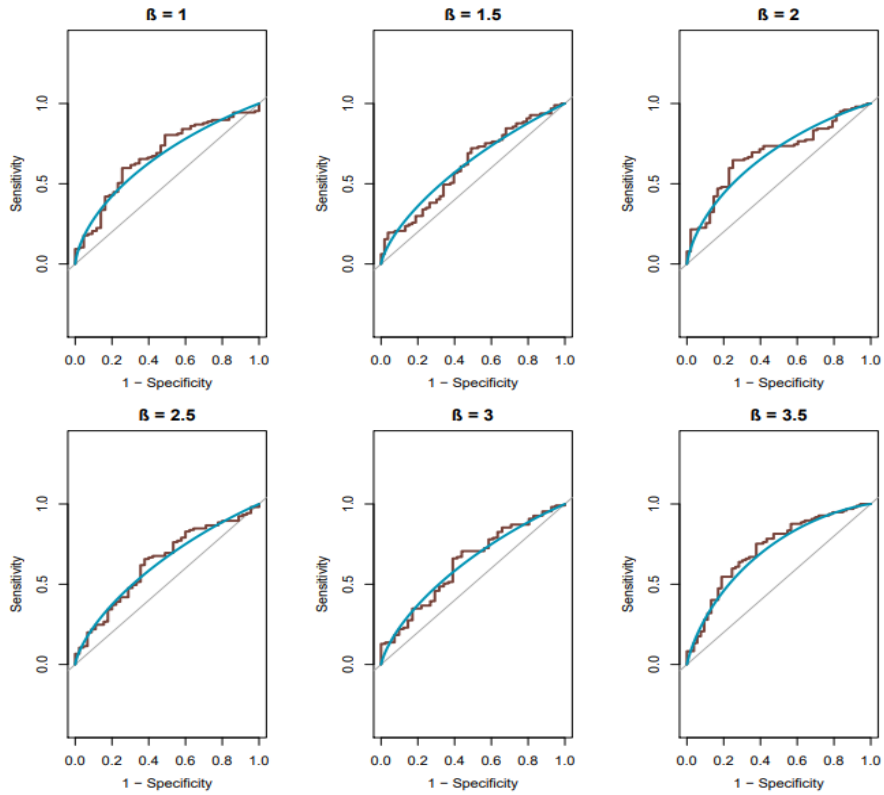
Rysunek 4.12: Estymowana oraz teoretyczna wartość krzywych ROC dla różnych parametrów α, oraz stałej wartości parametru $\beta = 1$ dla próby $n = 70$ - opracowanie własne



Rysunek 4.13: Estymowana oraz teoretyczna wartość krzywych ROC dla różnych parametrów β, oraz stałej wartości parametru $\alpha = 0.5$ dla próby $n = 70$ - opracowanie własne



Rysunek 4.14: Estymowana oraz teoretyczna wartość krzywych ROC dla różnych parametrów α, oraz stałej wartości parametru $\beta = 1$ dla próby $n = 150$ - opracowanie własne



Rysunek 4.15: Estymowana oraz teoretyczna wartość krzywych ROC dla różnych parametrów β, oraz stałej wartości parametru $\alpha = 0.5$ dla próby $n = 150$ - opracowanie własne

Z rysunków wyraźnie wynika, że dla małych próbek $n = 30$ dopasowanie empirycznych i teoretycznych krzywych ROC nie jest zbyt dokładne, zarówno w przypadku zmiennego parametru α i stałego β, jak i odwrotnym.

Przy zwiększeniu naszej próbki do $n = 70$, zauważamy, że estymowana oraz teoretyczna wartość krzywych dla różnych parametrów α oraz stałego parametru β nadal nie jest dokładna, ale w odwrotnej sytuacji, dla różnych parametrów β i stałego α możemy zauważyć już dość dobre dopasowanie. Dochodzimy do wniosku, że na dopasowanie krzywych teoretycznych i empirycznych I/D ROC większy wpływ ma niezmienność parametru α.

Dla już dość dużej próbki, $n = 150$ w obydwu przypadkach widzimy dobrze pokrywające się krzywe empiryczne i teoretyczne.

4.3 Analiza przeżycia na podstawie danych z biblioteki `survival` w `R`

Pakiet `survival` jest najczęściej używanym pakietem do analizy przeżycia w programie `R`. Zawiera on podstawowe procedury analizy przeżycia w tym na przykład krzywe Kaplana - Meiera czy modele Coxa. W tym rozdziale przeprowadzimy analizę danych `lung`, z wyżej wymienionej biblioteki, zawierających informacje na temat przeżycia pacjentów z zaawansowanym rakiem płuc z North Central Cancer Treatment Group. Naszym celem jest opisanie podstawowych koncepcji analizy przeżycia. W badaniach nad nowotworami większość analiz przeżycia wykorzystuje następujące metody, na których się skupimy:

- Wykresy Kaplana - Meiera do wizualizacji krzywych przeżycia
- Test log - rank do porównania krzywych przeżycia co najmniej dwóch grup
- Regresja proporcjonalnych hazardów Coxa w celu opisania wpływu zmiennych na przeżycie

Nasz zbiór zawiera 228 informacji o pacjentach, których przegląd charakterystyk obejmuje 10 następujących zmiennych:

- *inst* - kod instytucji
- *time* - czas przeżycia w dniach
- *status* - czy otrzymaliśmy zdarzenie (2 - śmierć), czy też nasze dane zostały ocenzurowane (1 - cenzura)
- *age* - wiek w latach
- *sex* - płeć, gdzie 1 oznacza mężczyznę, zaś 2 - kobietę
- *ph.ecog* - skala sprawności ECPG według lekarza (0 oznacza sprawność prawidłową, 5 - zgon)
- *ph.karno* - skala sprawności Karnofsky'ego według lekarza (100 oznacza sprawność prawidłową, 0 - zgon)

- *pat.karno* - skala sprawności Karnofsky'ego według pacjenta
- *meal.cal* - kalorie spożywane podczas posiłków
- *wt.loss* - utrata masy ciała w ciągu ostatnich sześciu miesięcy

Mamy zatem do czynienia zarówno ze zmiennymi jakościowymi jak i ilościowymi. Dane dotyczą zgonów pacjentów z zaawansowanym rakiem płuc w określonym przedziale czasowym. Część z obserwowanych osób nie umarła w całym okresie badania, przez co nasz zbiór posiada dane cenzurowane.

Pierwszą wykonaną przez nas czynnością, będzie przygotowanie danych do dalszej analizy.

```
library(survival)
pleć <- lung$sex
wiek <- lung$age
status <- lung$status
czas <- lung$time
```

Rysunek 4.16: Przygotowanie danych do dalszej analizy - opracowanie własne

Tabela poniżej przedstawia zestawienie ilości danych uciętych oraz otrzymanych zdarzeń.

Płeć	Ocenzurowanie	Śmierć	Suma osób
Mężczyzna	26	112	138
Kobieta	37	53	90

Tabela 4.2: Zestawienie danych dla zmiennych wiek i status - opracowanie własne

Zbiór zawiera 228 obserwacji, z czego aż $72,4\%$ pacjentów otrzymało zdarzenie. Widać wyraźną dominację danych cenzurowanych w 90 osobowej grupie kobiet (41%) niż w 138 osobowej grupie mężczyzn (19%). Może to być spowodowane zaniechaniem dalszych badań, nagminnym wypisywaniem się kobiet ze szpitala czy też śmiercią z innego powodu niż sprecyzowanego przez nas.

Korzystając z funkcji `Surv` oraz `survfit` zbadamy prawdopodobieństwo przeżycia względem płci. Wykorzystamy w tym miejscu dwa estymatory funkcji przeżycia - Kaplana - Meiera oraz Fleminga - Harringtona.

```
fit_KM <- survfit(Surv(czas, status) ~ pleć, data = lung,
                  type = "kaplan-meier")
fit_FH <- survfit(Surv(czas, status) ~ pleć, data = lung,
                  type = "fleming-harrington")
```

Rysunek 4.17: Funkcja badająca prawdopodobieństwo przeżycia względem płci - opracowanie własne

Płeć	Liczba obserwacji	Ilość zdarzeń	Mediana przeżycia	LCL	UCL
Mężczyźni	138	112	270	212	310
Kobiety	90	53	426	348	550

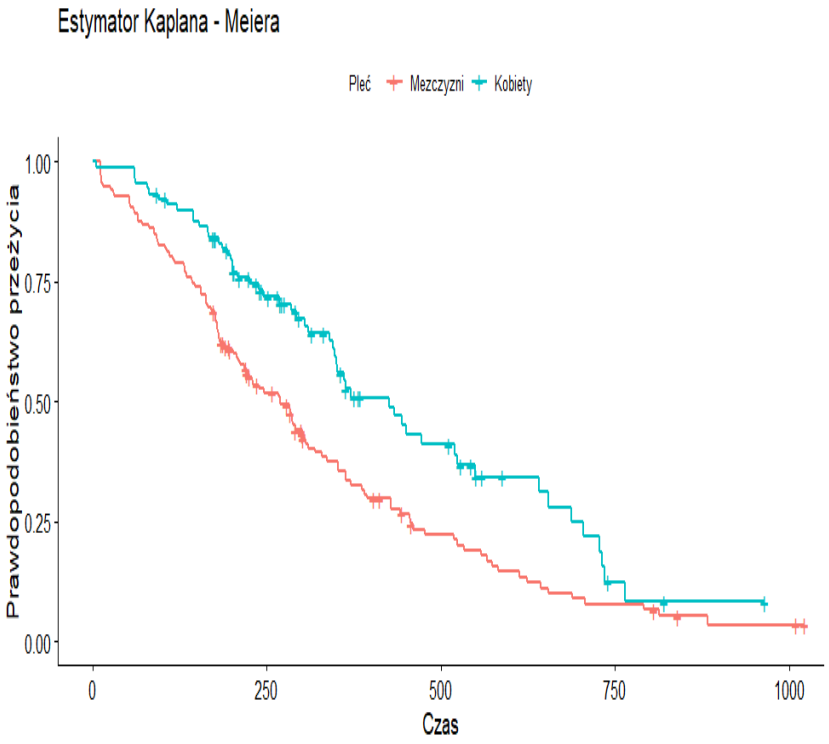
Tabela 4.3: Krótkie podsumowanie krzywych przeżycia estymatora Kaplana - Meiera - opracowanie własne

Płeć	Liczba obserwacji	Ilość zdarzeń	Mediana przeżycia	LCL	UCL
Mężczyźni	138	112	270	218	320
Kobiety	90	53	426	348	550

Tabela 4.4: Krótkie podsumowanie krzywych przeżycia estymatora Fleminga - Harringtona - opracowanie własne

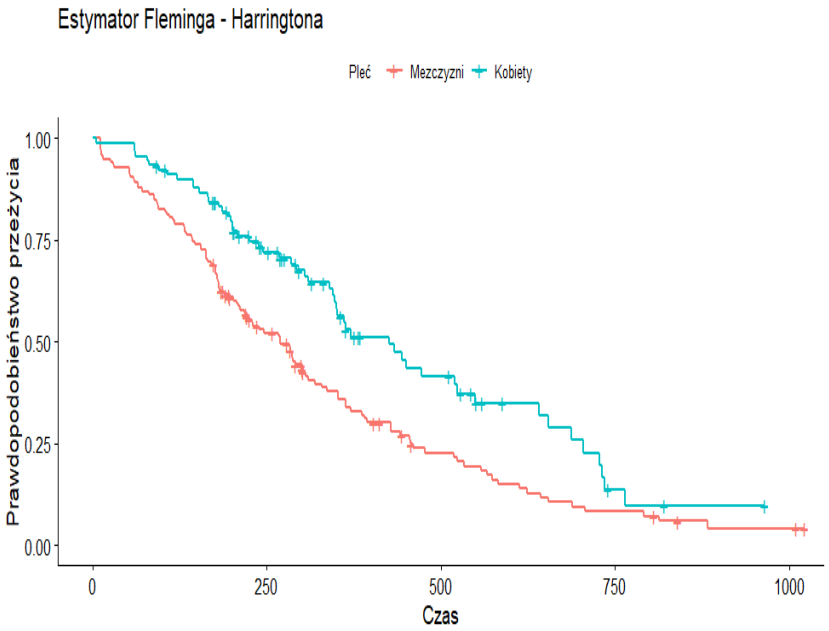
Mediana przeżycia kobiet jest prawie dwa razy większa niż mediana przeżycia mężczyzn. Sugeruje to, że większa śmiertelność z powodu raka płuc występuje u mężczyzn. Faktycznie, według Profesor Lucyny Mastalerz[1] rak płuc stanowi przyczynę 1/3 wszystkich zgonów z powodu nowotworów złośliwych u mężczyzn i 15% u kobiet.

Poniższe wykresy stanowią odzwierciedlenie kolejno tabel (4.3) oraz (4.4). Dokładnie widzimy, że potwierdzają one nasze wcześniejsze wnioski. Niebieska krzywa będąca w obydwu przypadkach estymatorem płci żeńskiej, przez cały okres badania znajduje się ponad czerwoną krzywą estymator przeżycia płci przeciwnej. To znaczy, że w całym okresie badania prawdopodobieństwo przeżycia kobiet jest większe.



Rysunek 4.18: Estymator Kaplana - Meiera funkcji płci - opracowanie własne

[1]Więcej informacji dostępnych pod adresem https://www.mp.pl/pacjent/onkologia/wywiady/99809,dlaczego-coraz-wiecej-kobiet-umiera-na-raka-pluc

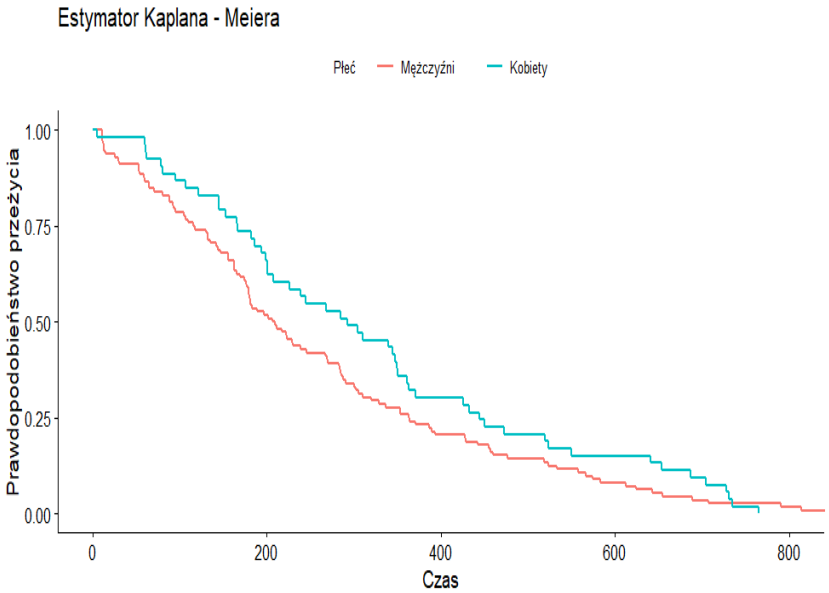


Rysunek 4.19: Estymator Fleminga - Harringtona funkcji płci - opracowanie własne

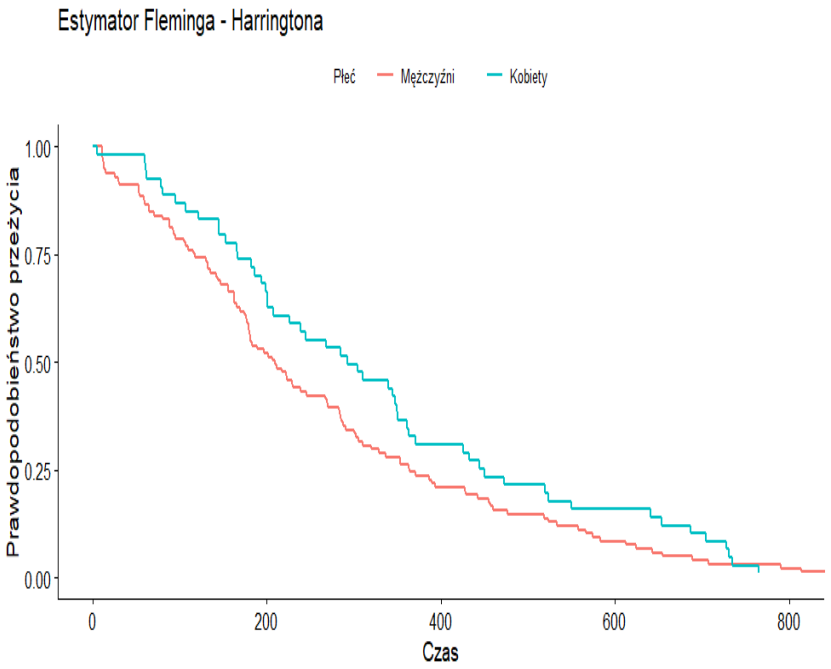
Jak już wcześniej zostało wspomniane, wiele danych dotyczących przeżycia kobiet zostało ocenzurowanych. Przefiltrujemy zatem zmienną status w celu utworzenia wektora zawierającego dane dotyczące otrzymania badanego zdarzenia. Innymi słowy, ze zmiennej status usuniemy wszystkie dane ocenzurowane.

```
died <- lung %>% filter(lung$status == 2)
```

Rysunek 4.20: Funkcja służąca do przefiltrowania zmiennej status względem otrzymania zdażenia - opracowanie własne



Rysunek 4.21: Estymator Kaplana - Meiera funkcji płci dla wszystkich, którzy otrzymali zdarzenie - opracowanie własne



Rysunek 4.22: Estymator Fleminga - Harringtona funkcji płci dla wszystkich, którzy otrzymali zdarzenie - opracowanie własne

Przeprowadzone analizy ukazały, że w każdym przypadku przeżycie kobiet jest większe, niż przeżycie mężczyzn.

Intuicyjnie mogłoby się wydawać, że wiek pacjenta może być ważnym czynnikiem różnicującym w kontekście umieralności na raka płuc. W naszym zbiorze danych znajduje się zmienna *age*, oznaczająca wiek w momencie przyjęcia pacjenta do szpitala. Chcąc zobaczyć wygląd estymatora Kaplana - Meiera w różnym przedziale wiekowym, wprowadzimy następującą kategoryzację zmiennej wiek (identyczną dla obu płci):

- pacjenci do 50 lat
- pacjenci między 50 a 65 rokiem życia
- pacjenci po 65 roku życia

Kategoria wiekowa	Liczba
1 - 50	26
50 - 65	110
65 - 100	92

Tabela 4.5: Kategoryzacja zmiennej wiek - opracowanie własne

Według Krajowego Rejestru Nowotworów[2] większość zachorowań na nowotwory złośliwe płuca występuje po 50 roku życia (96% zachorowań u mężczyzn, 95% zachorowań u kobiet), przy czym około 50% zachorowań u obu płci przypada powyżej 65 roku życia. W przypadku naszych danych `lung` proporcja ta również jest zachowana.

[2]http://onkologia.org.pl/nowotwory-zlosliwe-oplucnej-pluca-c33-34/

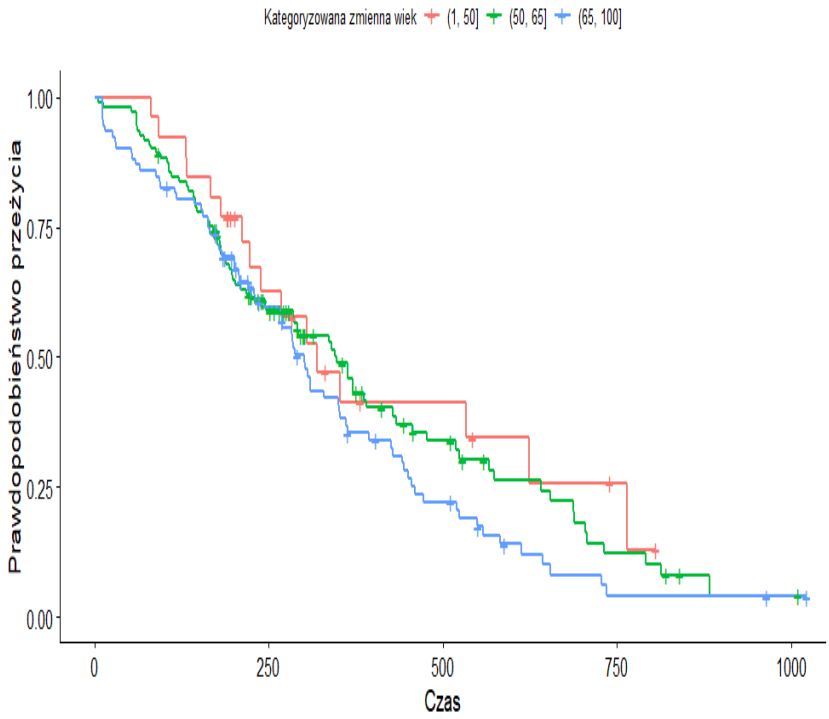
```
wiek <- cut(wiek, c(1, 50, 65, 100))
fit_wiek <- survfit(Surv(czas, status) ~ wiek, data = lung,
                    type = "kaplan-meier")
```

Rysunek 4.23: Funkcja analizująca przeżycie pacjentów według kategoryzowanej zmiennej wiek - opracowanie własne

Przedział	Liczba obserwacji	Ilość zdarzeń	Procentowa wartość	Mediana
$(0, 50]$	26	16	61%	320
$[50, 65)$	110	76	69%	348
$[65, 100]$	92	73	79%	301

Tabela 4.6: Krótkie podsumowanie kategoryzowanej zmiennej wiek - opracowanie własne

Tak jak przypuszczaliśmy, wraz z wiekiem szanse na przeżycie pacjenta chorującego na nowotwór złośliwy płuc maleją. Jednakże nawet chorując w młodym wieku, odsetek przeżyć nie jest zbyt duży. Możemy sobie zatem zadać pytanie, czy ten nowotwór może być jednym z najgroźniejszych? Jak podaje Światowa Organizacja Zdrowia (WHO) w 2018 roku rak płuc zabił 1,76 miliona osób na całym świecie, stając się tym samym najczęstszą przyczyną zgonów na nowotwór złośliwy (następny był rak jelita grubego, który zabił ponad dwa razy mniej ludzi - 862 000).



Rysunek 4.24: Estymator Kaplana - Meiera funkcji przeżycia kategoryzowanej zmiennej wiek - opracowanie własne

Powyżej przedstawiony został estymator funkcji przeżycia Kaplana - Meiera dla pacjentów z rakiem płuc z podziałem na kategoryzowaną zmienną wiek. Widzimy znacznie spadającą

funkcję schodkową, co potwierdza nasze wcześniejsze wnioski, że bez względu na wiek pacjenta, szanse na przeżycie są małe.

Funkcja `survdiff` testu log - rank testuje, czy istnieje różnica między dwiema (lub więcej) krzywymi przeżycia. Posiada ona parametr *rho* z przedziału $[0, 1]$, gdzie dla $rho = 0$ `survdiff` jest zwykłym testem log - rank zaś gdy $rho = 1$ mamy do czynienia z odpowiednikiem modyfikacji Peto & Peto testu Gehana - Wilcoxona. Na poziomie istotności $\alpha = 0.05$ rozpatrzymy teraz hipotezę zerową, o równości krzywych przeżycia dla obu płci, przeciwko hipotezie alternatywnej mówiącej o ich nierówności.

```
surv_plec <- survdiff(Surv(czas, status) ~ plec, data = lung)
```

Rysunek 4.25: Funkcja generująca test log - rank dla zmiennej płci - opracowanie własne

```
## Call:
## survdiff(formula = Surv(czas, status) ~ plec, data = lung)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## plec=1 138      112     91.6      4.55      10.3
## plec=2  90       53     73.4      5.68      10.3
##
##  Chisq= 10.3  on 1 degrees of freedom, p= 0.001
```

Rysunek 4.26: Test log - rank dla zmiennej płci - opracowanie własne

W tym przypadku $p - value = 0.001$ jest mniejsza od poziomu istotności $\alpha = 0.05$, zatem odrzucamy hipotezę zerową o równości krzywych przeżycia dla pacjentów obu płci. Następnie w celu sprawdzenia wartości statystyk podniesiemy wartość *rho* do 1 (to znaczy, większą wagę przyłożymy do zdarzeń wcześniejszych).

```
## Call:
## survdiff(formula = Surv(czas, status) ~ plec, data = lung, rho = 1)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## plec=1 138     70.4     55.6      3.95      12.7
## plec=2  90     28.7     43.5      5.04      12.7
##
##  Chisq= 12.7  on 1 degrees of freedom, p= 4e-04
```

Rysunek 4.27: Funkcja generująca test Peto & Peto testu Gehana - Wilcoxona dla zmiennej płci - opracowanie własne

Tym razem wartość *p-value* jest rzędu 10^{-4}, zatem ponownie zdecydowanie odrzucamy naszą hipotezę zerową. Dodatkowo przyglądając się pozostałym komponentom w obydwu powyższych tabelach, widzimy, że ważona oczekiwana liczba zdarzeń (*Expected*) jest zawsze wyższa u mężczyzn, niż kobiet. Stąd wniosek, że umieralność na nowotwór płuc jest większa dla pacjentów płci męskiej niż płci żeńskiej.

Estymator funkcji przeżycia dla kategoryzowanej zmiennej wiek dał nam wniosek, że śmiertelność nie jest widocznie zależną funkcją od wieku w chwili zachorowania. Korzystając ponownie z funkcji `survdiff` oraz testu log - rank sprawdzimy *p-value* dla wyznaczonych wcześniej trzech poziomów wieku.

```
surv_wiek <- survdiff(Surv(czas, status) ~ wiek, data = lung)
```

Rysunek 4.28: Funkcja generująca test log - rank dla kategoryzowanej zmiennej wiek - opracowanie własne

```
## Call:
## survdiff(formula = Surv(czas, status) ~ wiek, data = lung)
##
##                 N Observed Expected (O-E)^2/E (O-E)^2/V
## wiek=(1,50]    26       16     20.6     1.019     1.175
## wiek=(50,65]  110       76     81.9     0.424     0.847
## wiek=(65,100]  92       73     62.5     1.753     2.855
##
##  Chisq= 3.2  on 2 degrees of freedom, p= 0.2
```

Rysunek 4.29: Test log - rank dla kategoryzowanej zmiennej wiek - opracowanie własne

Wartość $p-value = 0.2 > 0.05 = \alpha$, zatem nie ma podstaw do odrzucenia hipotezy zerowej. Przyjmujemy zatem, że większa śmiertelność faktycznie nie zależy od wcześniejszego czy późniejszego wieku.

Skorzystamy jeszcze z modyfikacji Peto & Peto testu Gehana - Wilcoxona:

```
## Call:
## survdiff(formula = Surv(czas, status) ~ wiek, data = lung, rho = 1)
##
##                 N Observed Expected (O-E)^2/E (O-E)^2/V
## wiek=(1,50]    26     9.45     12.1     0.563     0.907
## wiek=(50,65]  110    45.41     48.4     0.187     0.514
## wiek=(65,100]  92    44.25     38.6     0.815     1.877
##
##  Chisq= 2.2  on 2 degrees of freedom, p= 0.3
```

Rysunek 4.30: Funkcja generująca test Peto & Peto testu Gehana - Wilcoxona dla kategoryzowanej zmiennej wiek - opracowanie własne

Ponownie, $p - value = 0.3 > 0.05 = \alpha$, zatem nie mamy podstaw do odrzucenia hipotezy zerowej.

Korzystając z nieparametrycznego modelu Coxa, wykonamy dwa testy.

1. *I* test dotyczący zmiennej płci
   - H_0 = równość funkcji przeżycia ze względu na płeć
   - H_1 = brak równości funkcji przeżycia ze względu na płeć

2. II test dotyczący zmiennej wiek
   - H_0 = równość funkcji przeżycia ze względu na wiek
   - H_1 = brak równości funkcji przeżycia ze względu na wiek

```
## Call:
## coxph(formula = Surv(czas, status) ~ plec, data = lung)
##
##   n= 228, number of events= 165
##
##         coef exp(coef) se(coef)      z Pr(>|z|)
## plec -0.5310    0.5880   0.1672 -3.176  0.00149 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      exp(coef) exp(-coef) lower .95 upper .95
## plec     0.588      1.701    0.4237     0.816
##
## Concordance= 0.579  (se = 0.021 )
## Likelihood ratio test= 10.63  on 1 df,   p=0.001
## Wald test            = 10.09  on 1 df,   p=0.001
## Score (logrank) test = 10.33  on 1 df,   p=0.001
```

Rysunek 4.31: Test Coxa dla zmiennej płci - opracowanie własne

Jak wynika z powyższych danych, $p - value = 0.001 < 0.05 = \alpha$, zatem odrzucamy hipotezę zerową i przyjmujemy hipotezę alternatywną, mówiącą o braku równości funkcji przeżycia ze względu na płeć badanego.

```
## Call:
## coxph(formula = Surv(czas, status) ~ age, data = lung)
##
##   n= 228, number of events= 165
##
##         coef exp(coef) se(coef)     z Pr(>|z|)
## age 0.018720  1.018897 0.009199 2.035   0.0419 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     exp(coef) exp(-coef) lower .95 upper .95
## age     1.019     0.9815     1.001     1.037
##
## Concordance= 0.55  (se = 0.025 )
## Likelihood ratio test= 4.24  on 1 df,   p=0.04
## Wald test            = 4.14  on 1 df,   p=0.04
## Score (logrank) test = 4.15  on 1 df,   p=0.04
```

Rysunek 4.32: Test Coxa dla całej zmiennej wiek - opracowanie własne

Powyższe wyniki świadczą o tym, że nie mamy podstaw do odrzucenia hipotezy zerowej ($p - value = 0.04 > 0.05 = \alpha$), zakładamy więc, że występuje równości funkcji przeżycia ze względu na całą zmienną wiek.

Zbadamy jeszcze, czy do takiego samego wniosku dojdziemy dla kategoryzowanej zmiennej wiek.

```
## Call:
## coxph(formula = Surv(czas, status) ~ wiek, data = lung)
##
##   n= 228, number of events= 165
##
##                coef exp(coef) se(coef)     z Pr(>|z|)
## wiek(50,65]  0.1791    1.1961   0.2756 0.650    0.516
## wiek(65,100] 0.4114    1.5089   0.2771 1.484    0.138
##
##              exp(coef) exp(-coef) lower .95 upper .95
## wiek(50,65]      1.196     0.8360    0.6969     2.053
## wiek(65,100]     1.509     0.6627    0.8765     2.597
##
## Concordance= 0.531  (se = 0.023 )
## Likelihood ratio test= 3.25  on 2 df,   p=0.2
## Wald test            = 3.21  on 2 df,   p=0.2
## Score (logrank) test = 3.24  on 2 df,   p=0.2
```

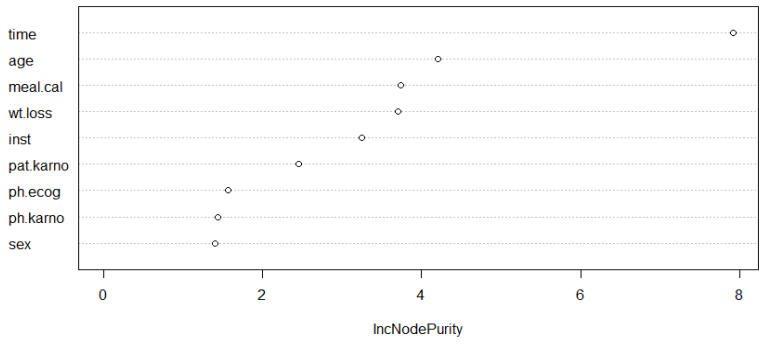
Rysunek 4.33: Test Coxa dla kategoryzowanej zmiennej wiek - opracowanie własne

Test Coxa pozwala wnioskować, że przeżycie badanej grupy pacjentów jest funkcją zależną od płci, ale nie zależy od wieku, bez względu czy analizujemy pełną grupę wiekową, czy też ją kategoryzujemy.

W dalszej części rozdziału zajmiemy się analizą krzywych ROC na różnych przedziałach czasowych. Zanim jednak do tego przejdziemy, zbadamy ranking ważności analizowanych zmiennych. Wykorzystamy w tym celu funkcję `varImpPlot` z biblioteki `randomForest`.

```
##              IncNodePurity
## inst          3.421083
## time          7.951373
## age           4.203949
## sex           1.394122
## ph.ecog       1.391022
## ph.karno      1.493526
## pat.karno     2.355440
## meal.cal      3.734905
## wt.loss       3.682298
```

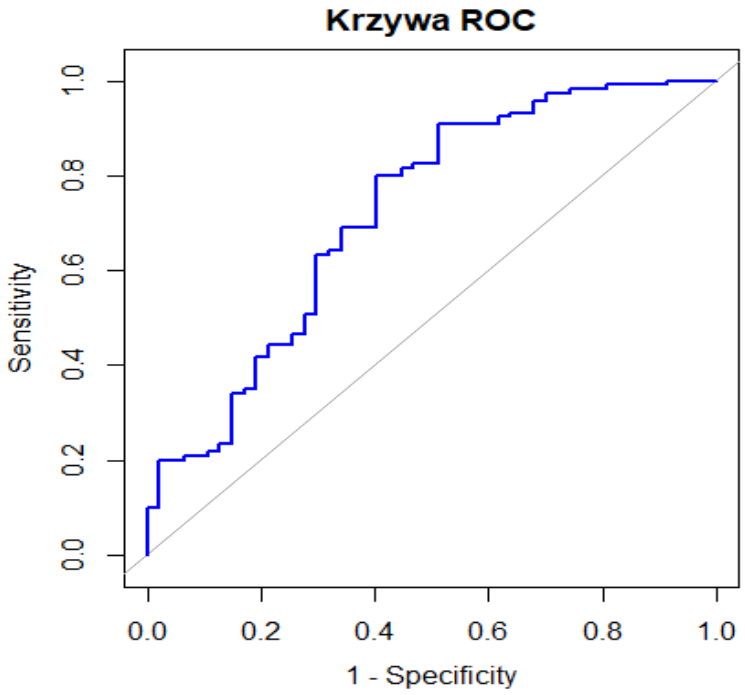
Rysunek 4.34: Ważność zmiennych losowych - opracowanie własne



Rysunek 4.35: Wykres ważności zmiennych losowych - opracowanie własne

Im wyższa wartość parametru *IncNodePurity*, tym większe znaczenie zmiennej w naszym modelu. Analizując powyższe rysunki, dochodzimy do wniosku, że największe znaczenie w naszym modelu mają dwie zmienne - *time* oraz *age*, zaś najmniejsze - *sex*, *ph.ecog* oraz *ph.karno*.

Korzystając teraz z biblioteki `pROC` oraz funkcji `roc` generujemy krzywą ROC na podstawie naszych danych lung.



Rysunek 4.36: Wykres krzywej ROC dla zadanej kombinacji zmiennych - opracowanie własne

STATUS
INST+TIME+AGE+SEX+PH.ECOG+PH.KARNO+PAT.KARNO+MEAL.CAL+WT.LOSS
0.7213

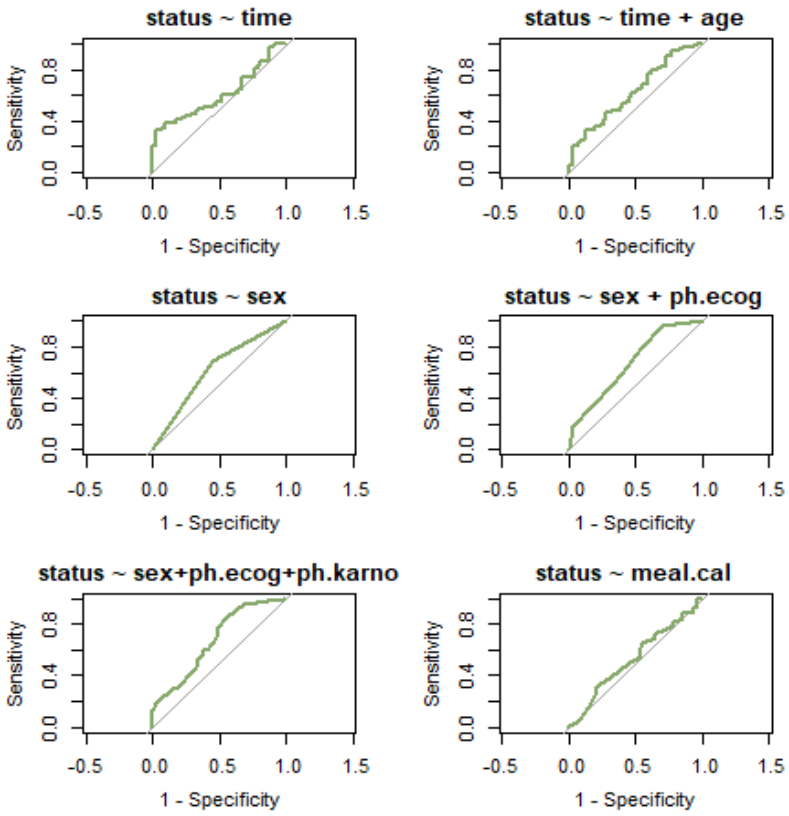
Tabela 4.7: Wartości pola AUC dla zadanej kombinacji zmiennych - opracowanie własne

STATUS	TIME	TIME + AGE	SEX	SEX + PH.ECOG
	0.6085	0.6259	0.6183	0.6788

Tabela 4.8: Wartości pola AUC dla zadanej kombinacji zmiennych - opracowanie własne

STATUS	SEX + PH.ECOG + PH.KARNO	MEAL.CAL
	0.6813	0.5373

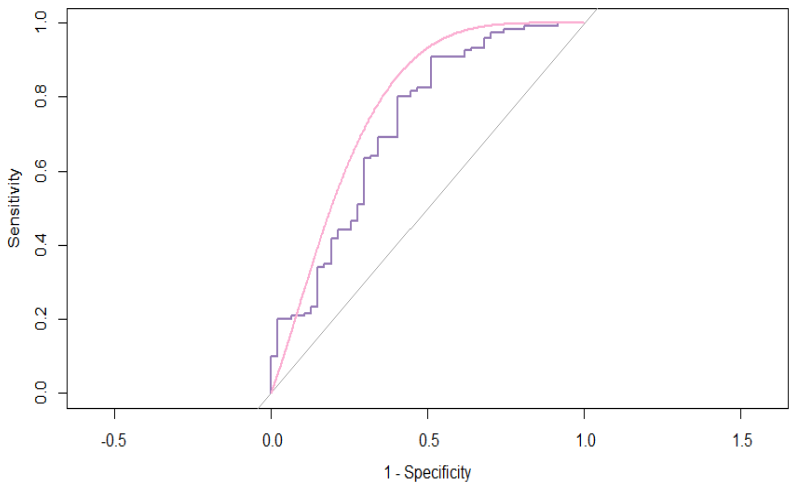
Tabela 4.9: Wartości pola AUC dla zadanej kombinacji zmiennych - opracowanie własne



Rysunek 4.37: Wykres krzywej ROC dla zadanej kombinacji zmiennych - opracowanie własne

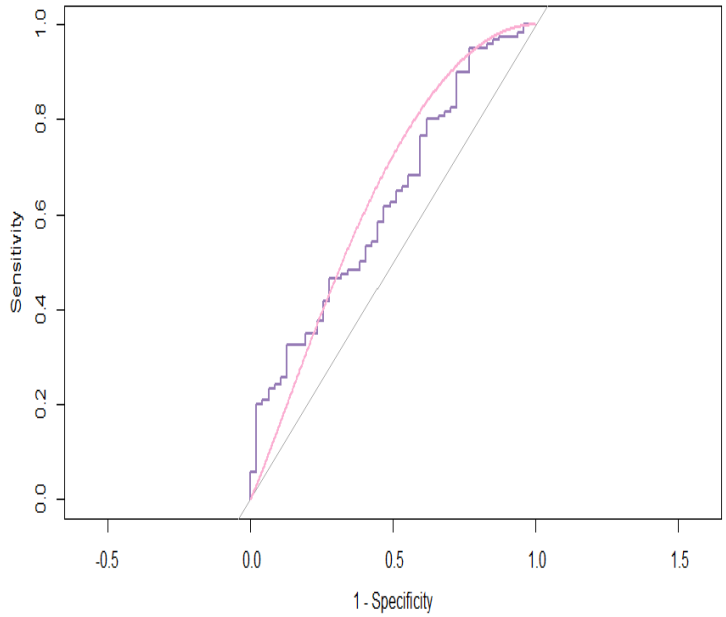
Tak jak wcześniej, za zmienną zależną przyjmujemy zmienną *status*, zaś za charakterystyki zmienne przyjmiemy trzy grupy, dla których kolejno sprawdzimy dopasowanie krzywych empirycznych oraz teoretycznych.

- $status \sim inst + time + status + age + sex + ph.ecog + ph.karno + pat.karno + meal.cal + wt.loss$



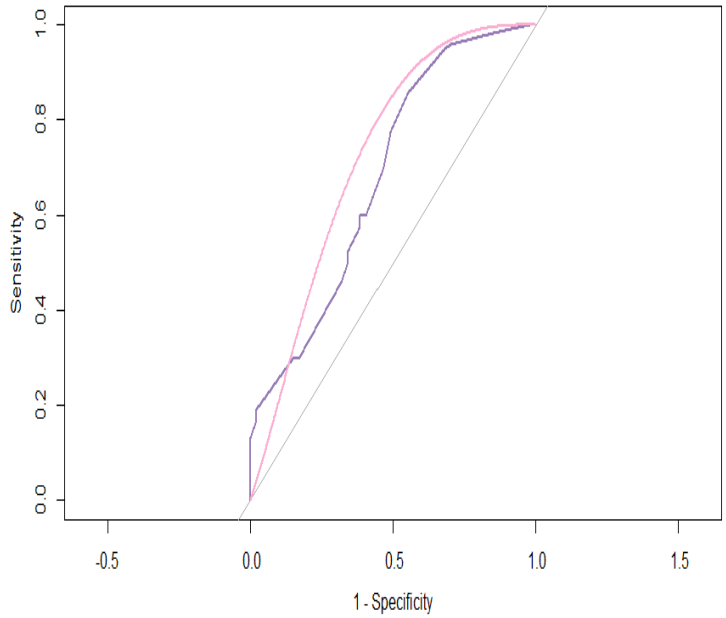
Rysunek 4.38: Wykres dopasowania empirycznej i teoretycznej krzywej ROC dla zadanej kombinacji zmiennych - opracowanie własne

- $status \sim time + age$



Rysunek 4.39: Wykres dopasowania empirycznej i teoretycznej krzywej ROC dla zadanej kombinacji zmiennych - opracowanie własne

- $status \sim sex + ph.ecog + ph.karno$



Rysunek 4.40: Wykres dopasowania empirycznej i teoretycznej krzywej ROC dla zadanej kombinacji zmiennych - opracowanie własne

Funkcja `timeROC` w bibliotece `timeROC` oszacowuje odwrotne prawdopodobieństwa cenzurowania skumulowanej/dynamicznej krzywej ROC zależnej od czasu. Jest ona następującej postaci:

```
timeROC(T, delta, marker, other_markers = NULL, cause,
            weighting = "marginal", times, ROC = TRUE, iid = FALSE)
```

Rysunek 4.41: Funkcja generująca krzywą ROC zależną od czasu - źródło[3]

Gdzie

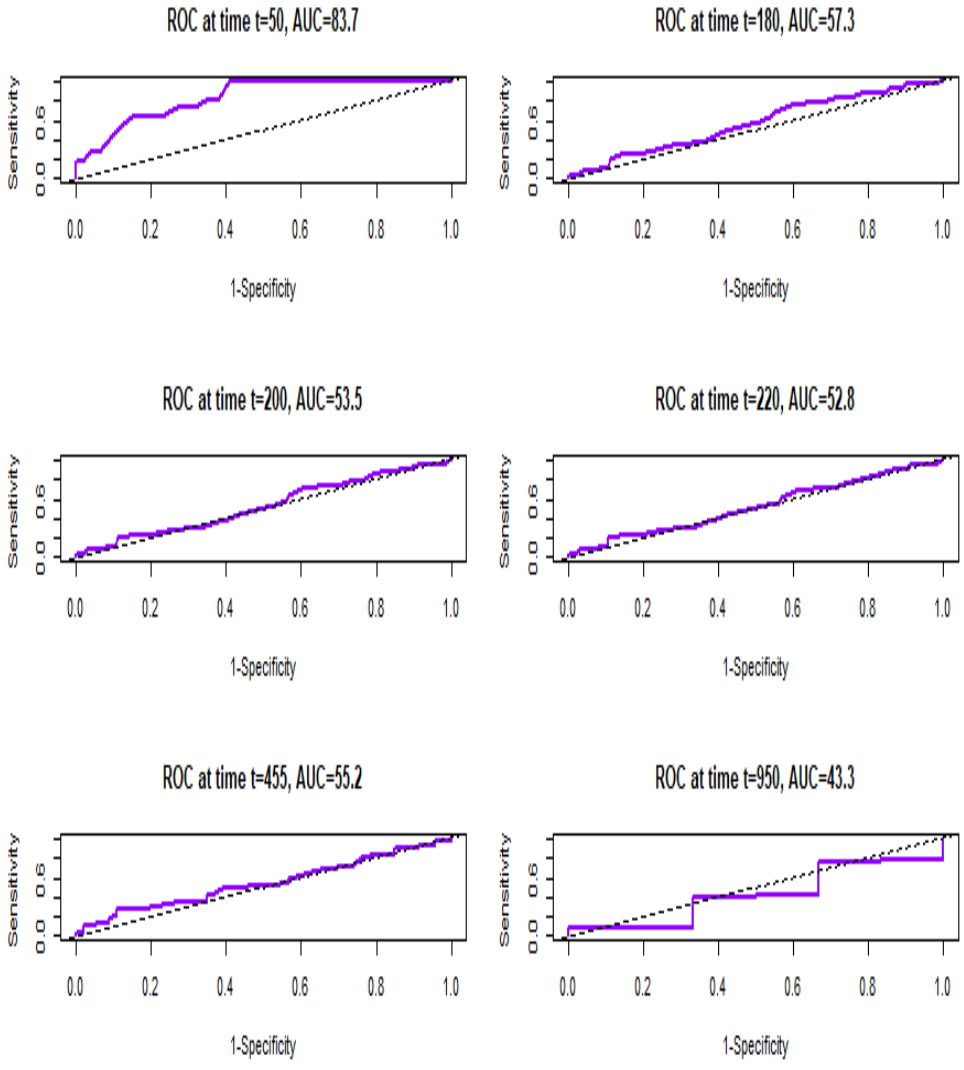
- T - wektor ocenzurowanych czasów zdarzeń
- *delta* - wektor wskaźników zdarzeń przy odpowiedniej wartości wektora T. Obserwacje ocenzurowane muszą być oznaczone wartością 0.
- *marker* - wektor wartości znaczników, dla których chcemy obliczyć zależne od czasu krzywe ROC. Co ważne, funkcja ta zakłada, że większe wartości znacznika wiążą się z wyższym ryzykiem zdarzeń.
- *othermarkers* - macierz zawierająca wartości innych znaczników, które chcemy uwzględnić przy obliczaniu odwrotnego prawdopodobieństwa cenzurowania wag.
- *cause* - wartość wskaźnika zdarzenia reprezentująca interesujące nas zdarzenie, dla którego zamierzamy obliczyć zależną od czasu krzywą ROC
- *weighting* - metoda używana do obliczania wag. Domyślna wartość to *weighting* = "*marginal*", która korzysta z estymatora Kaplana - Meiera rozkładu cenzurowania. Można także wybrać wartość *weighting* = "*cox*", bądź *weighting* = "" modelujące odpowiednio cenzurowanie za pomocą modelu Coxa i addytywnego modelu Aalen
- *times* - wektor czasów punktów t dla których chcemy obliczyć zależną od czasu krzywą ROC
- *ROC* - wartość logiczna wskazująca na to, czy chcemy zapisać oszacowania wrażliwości i swoistości
- *iid* - wartość logiczna wskazująca, czy chcemy obliczyć iid - reprezentację obszary pod estymatorem krzywej ROC zależnej od czasu

Korzystając najpierw z metody *weighting* = "*marginal*", a następnie *weighting* = "*cox*" ocenimy kolejno wiek, płeć, ph.ecog oraz ph.karno jako prognostyczny biomarker śmierci.

[3]https://rdrr.io/cran/timeROC/man/timeROC.html

```
ROC.age <- timeROC(T = lung$time, delta = lung$status,
                   weighting = "marginal",
                   marker = lung$age, cause = 1,
                   times=c(50, 180, 200, 220, 455, 950))
ROC.age

## Time-dependent-Roc curve estimated using IPCW  (n=227, without competing risks).
##       Cases Survivors Censored AUC (%)
## t=50     11       216        0   83.69
## t=180    62       158        7   57.32
## t=220    80       132       15   52.84
## t=455   132        46       49   55.17
## t=950   164         3       60   43.26
##
## Method used for estimating IPCW:marginal
##
## Total computation time : 0.01  secs.
```

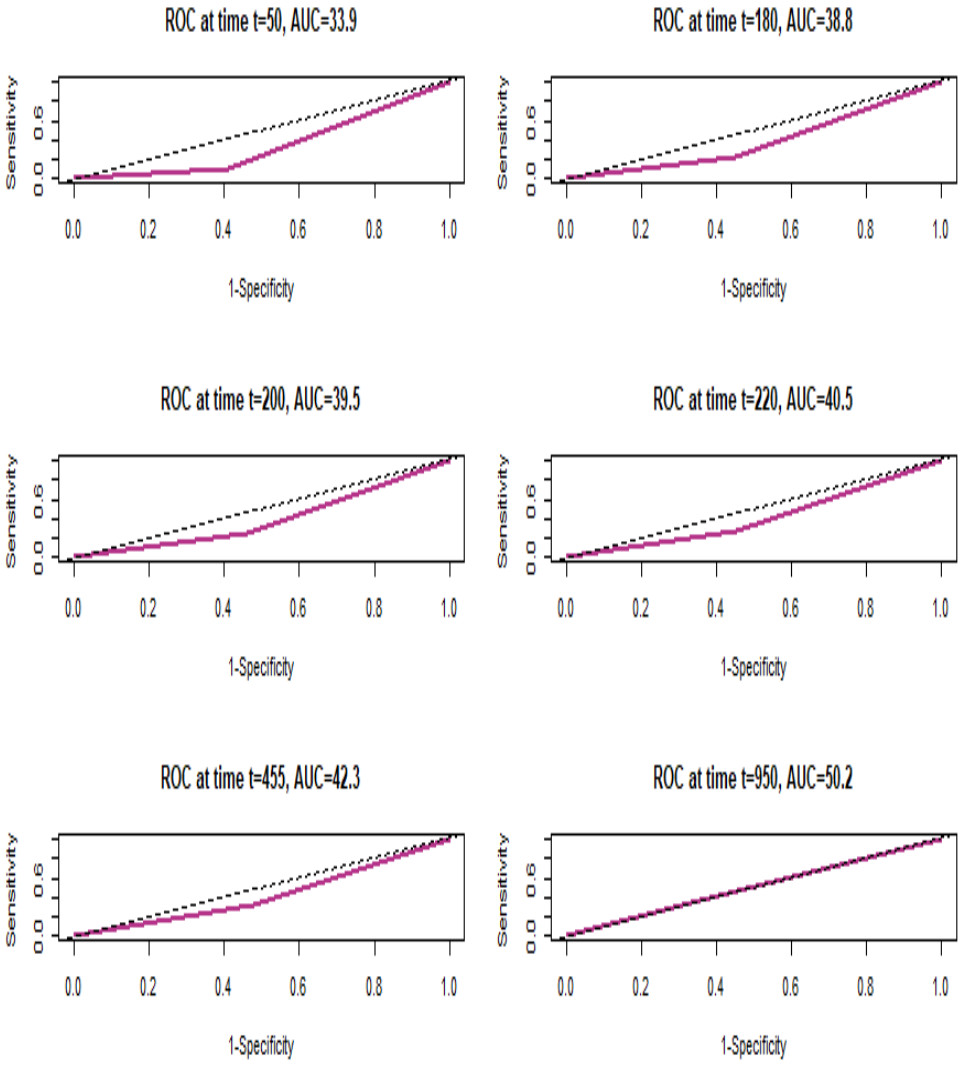
Rysunek 4.42: Time - dependent ROC metodą Kaplana - Meiera dla zmiennej wiek - opracowanie własne



Rysunek 4.43: Time - dependent ROC metodą Kaplana - Meiera dla zmiennej wiek - opracowanie własne

```
ROC.sex <- timeROC(T = lung$time, delta = lung$status,
                   weighting = "marginal",
                   marker = lung$sex, cause = 1,
                   times=c(50, 180, 200, 220, 455, 950))
ROC.sex

## Time-dependent-Roc curve estimated using IPCW  (n=227, without competing risks).
##       Cases Survivors Censored AUC (%)
## t=50     11       216        0   33.94
## t=180    62       158        7   38.81
## t=220    80       132       15   40.47
## t=455   132        46       49   42.26
## t=950   164         3       60   50.16
##
## Method used for estimating IPCW:marginal
##
## Total computation time : 0  secs.
```

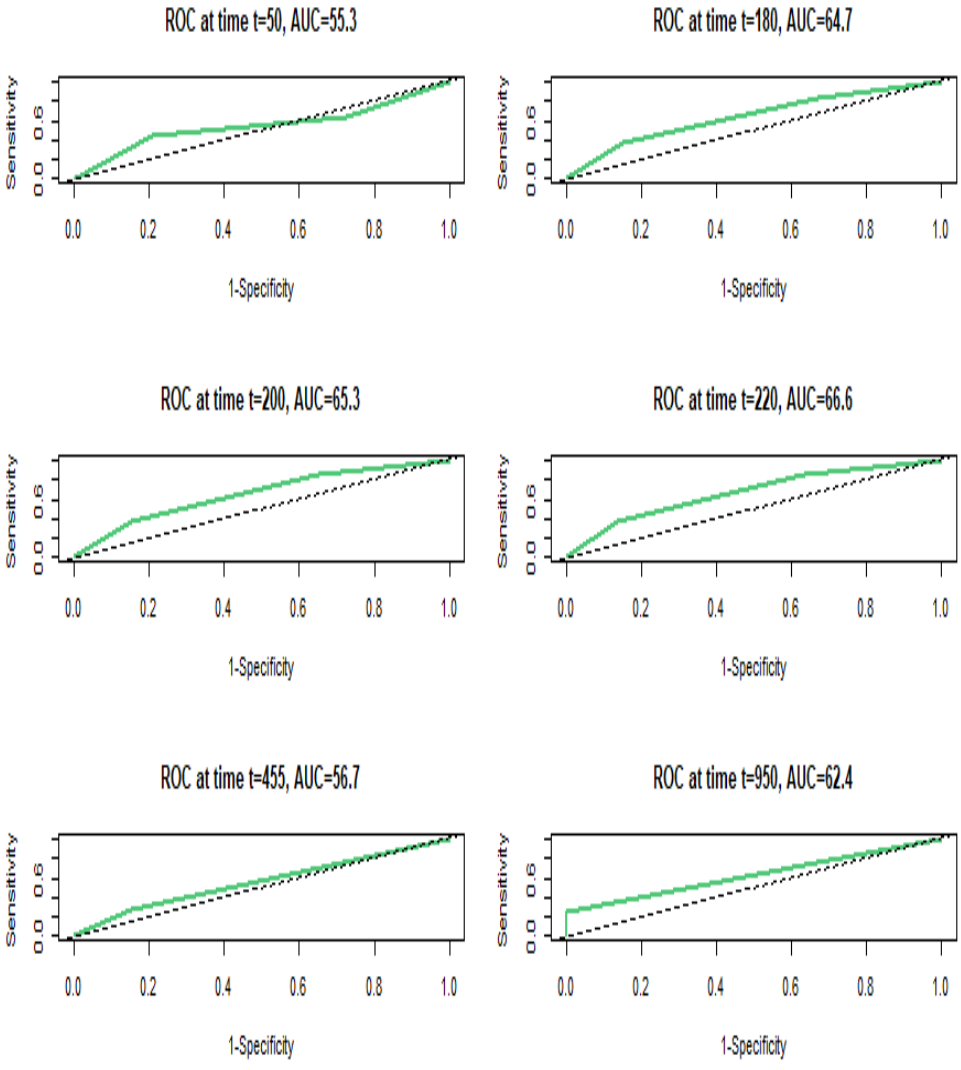
Rysunek 4.44: Time - dependent ROC metodą Kaplana - Meiera dla zmiennej płci - opracowanie własne



Rysunek 4.45: Time - dependent ROC metodą Kaplana - Meiera dla zmiennej płci - opracowanie własne

```
ROC.ph.ecog <- timeROC(T = lung$time, delta = lung$status,
                       weighting = "marginal",
                       marker = lung$ph.ecog, cause = 1,
                       times=c(50, 180, 200, 220, 455, 950))
ROC.ph.ecog

## Time-dependent-Roc curve estimated using IPCW  (n=226, without competing risks).
##        Cases Survivors Censored AUC (%)
## t=50      11       215        0   55.26
## t=180     61       158        7   64.68
## t=220     79       132       15   66.59
## t=455    131        46       49   56.69
## t=950    163         3       60   62.43
##
## Method used for estimating IPCW:marginal
##
## Total computation time : 0  secs.
```

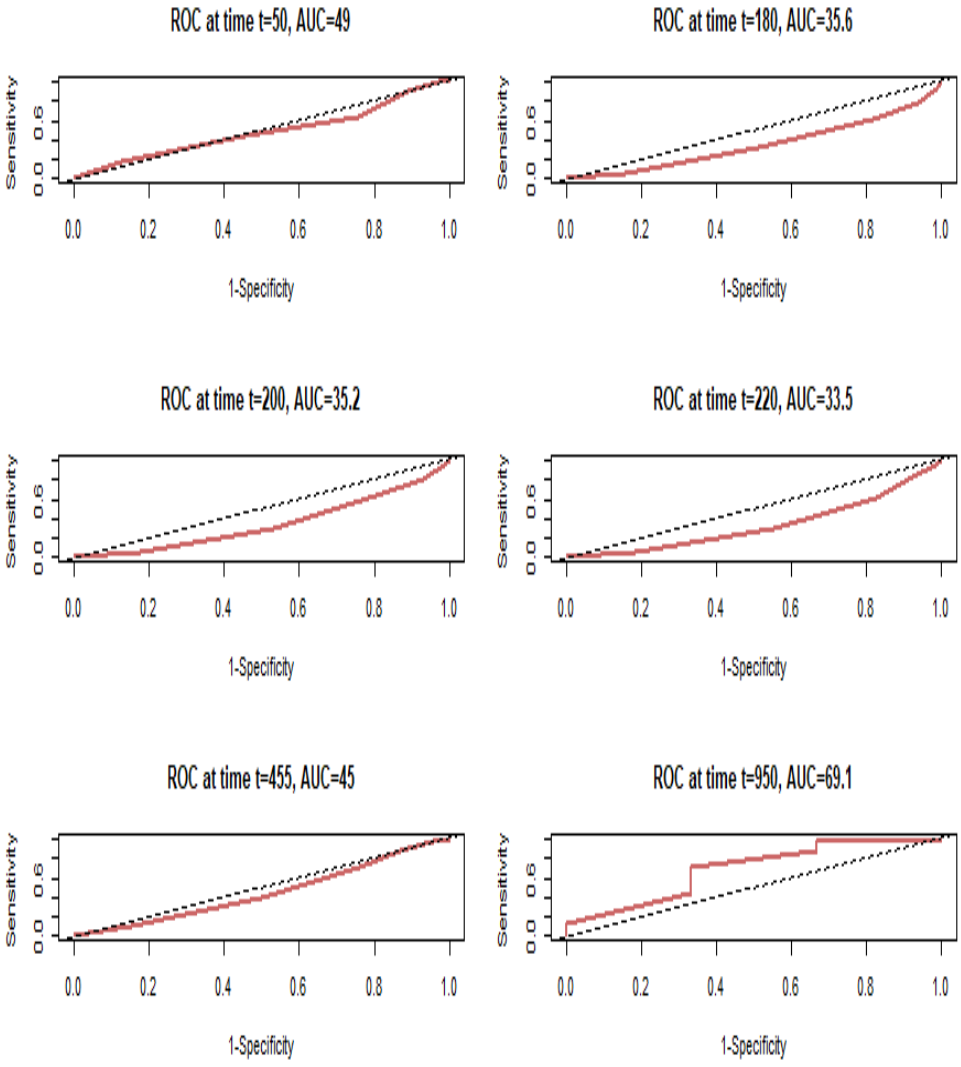
Rysunek 4.46: Time - dependent ROC metodą Kaplana - Meiera dla zmiennej ph.ecog - opracowanie własne



Rysunek 4.47: Time - dependent ROC metodą Kaplana - Meiera dla zmiennej ph.ecog - opracowanie własne

```
ROC.ph.karno <- timeROC(T = lung$time, delta = lung$status,
                        weighting = "marginal",
                        marker = lung$ph.karno, cause = 1,
                        times=c(50, 180, 200, 220, 455, 950))
ROC.ph.karno

## Time-dependent-Roc curve estimated using IPCW  (n=226, without competing risks).
##       Cases Survivors Censored AUC (%)
## t=50     11       215        0   48.99
## t=180    61       158        7   35.63
## t=220    79       132       15   33.54
## t=455   131        46       49   44.95
## t=950   163         3       60   69.05
##
## Method used for estimating IPCW:marginal
##
## Total computation time : 0  secs.
```

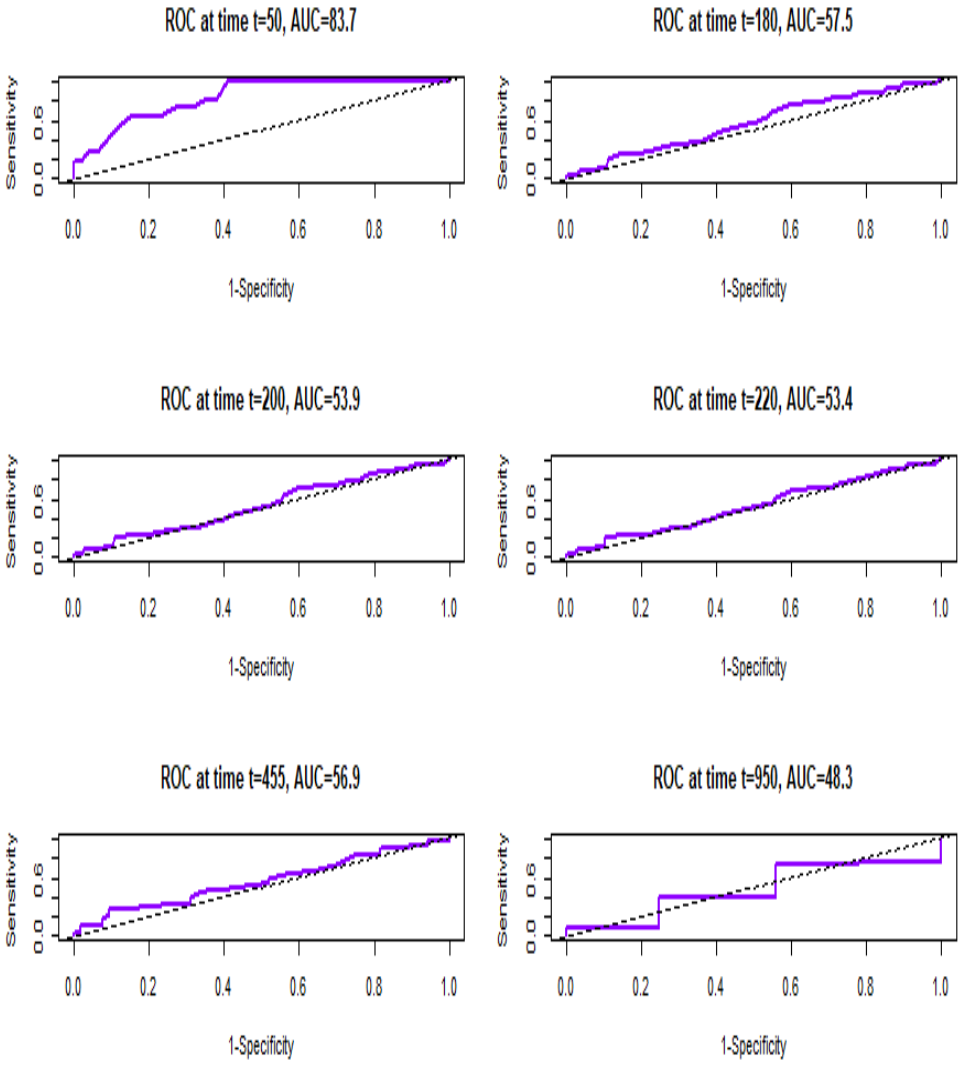
Rysunek 4.48: Time - dependent ROC metodą Kaplana - Meiera dla zmiennej ph.karno - opracowanie własne



Rysunek 4.49: Time - dependent ROC metodą Kaplana - Meiera dla zmiennej ph.karno - opracowanie własne

```
ROC.age <- timeROC(T = lung$time, delta = lung$status,
                   weighting = "cox",
                   marker = lung$age, cause = 1,
                   times=c(50, 180, 200, 220, 455, 950))
ROC.age

## Time-dependent-Roc curve estimated using IPCW  (n=227, without competing risks).
##       Cases Survivors Censored AUC (%)
## t=50     11       216        0   83.69
## t=180    62       158        7   57.53
## t=220    80       132       15   53.35
## t=455   132        46       49   56.85
## t=950   164         3       60   48.27
##
## Method used for estimating IPCW:cox
##
## Total computation time : 1.64  secs.
```

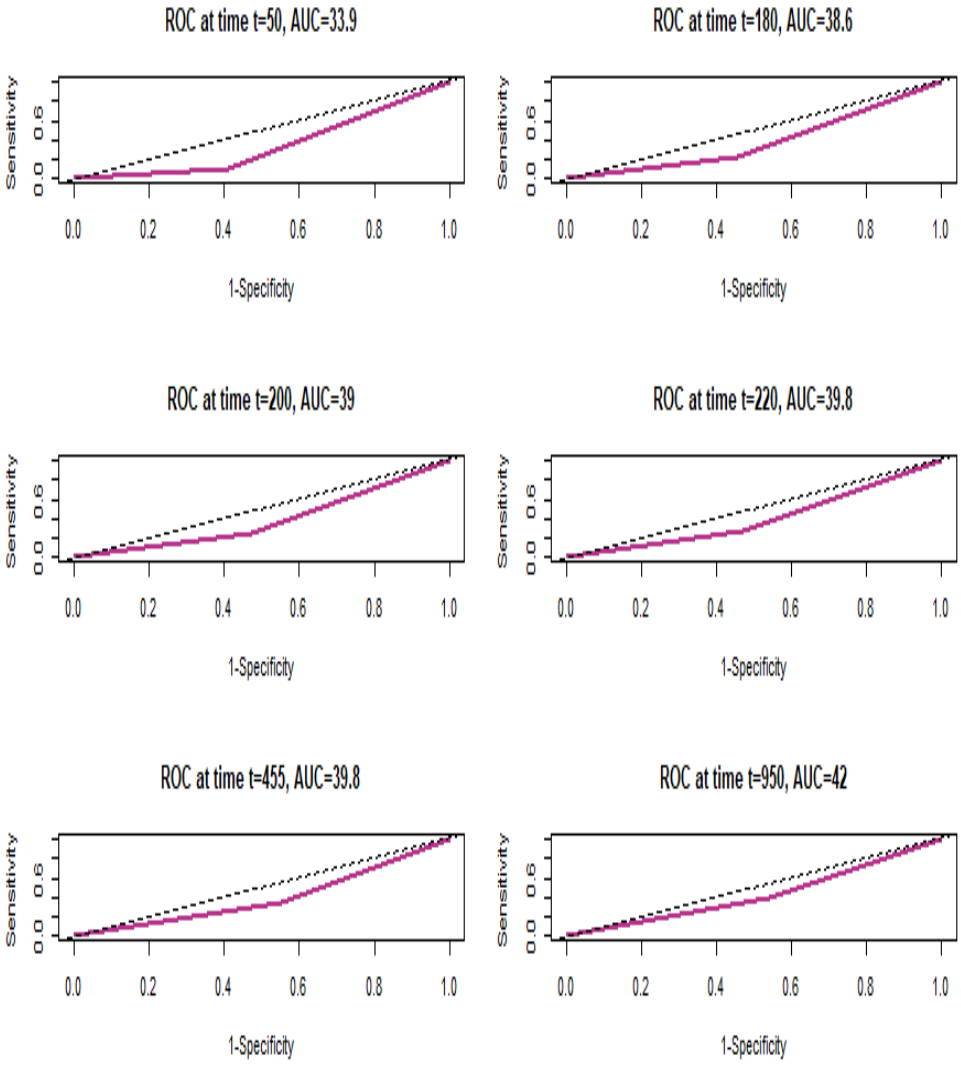
Rysunek 4.50: Time - dependent ROC metodą Coxa dla zmiennej wiek - opracowanie własne



Rysunek 4.51: Time - dependent ROC metodą Coxa dla zmiennej wiek - opracowanie własne

```
ROC.sex <- timeROC(T = lung$time, delta = lung$status,
                   weighting = "cox",
                   marker = lung$sex, cause = 1,
                   times=c(50, 180, 200, 220, 455, 950))
ROC.sex

## Time-dependent-Roc curve estimated using IPCW  (n=227, without competing risks).
##       Cases Survivors Censored AUC (%)
## t=50     11       216        0   33.94
## t=180    62       158        7   38.57
## t=220    80       132       15   39.85
## t=455   132        46       49   39.76
## t=950   164         3       60   42.05
##
## Method used for estimating IPCW:cox
##
## Total computation time : 0.01  secs.
```

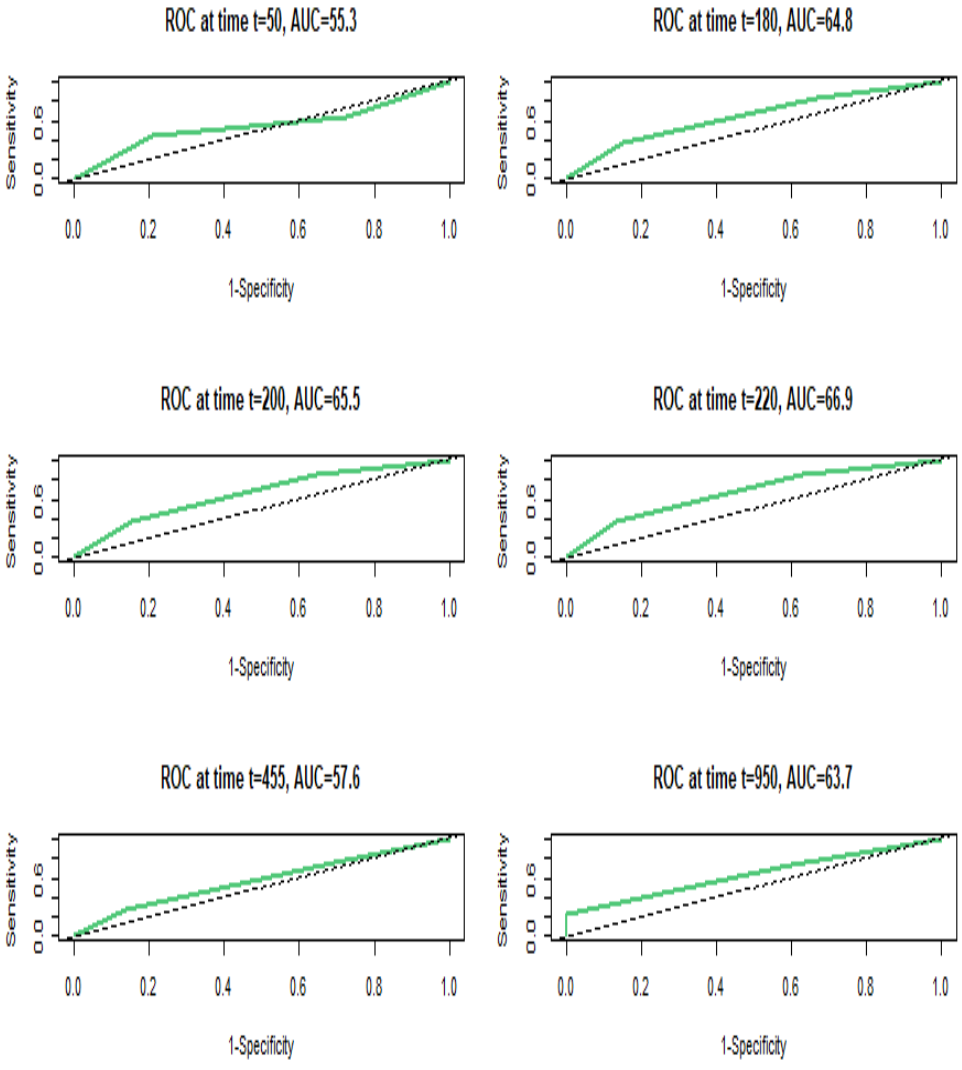
Rysunek 4.52: Time - dependent ROC metodą Coxa dla zmiennej płci - opracowanie własne



Rysunek 4.53: Time - dependent ROC metodą Coxa dla zmiennej płci - opracowanie własne

```
ROC.ph.ecog <- timeROC(T = lung$time, delta = lung$status,
                       weighting = "cox",
                       marker = lung$ph.ecog, cause = 1,
                       times=c(50, 180, 200, 220, 455, 950))
ROC.ph.ecog

## Time-dependent-Roc curve estimated using IPCW  (n=226, without competing risks).
##       Cases Survivors Censored AUC (%)
## t=50     11       215        0   55.26
## t=180    61       158        7   64.78
## t=220    79       132       15   66.86
## t=455   131        46       49   57.55
## t=950   163         3       60   63.71
##
## Method used for estimating IPCW:cox
##
## Total computation time : 0.02  secs.
```

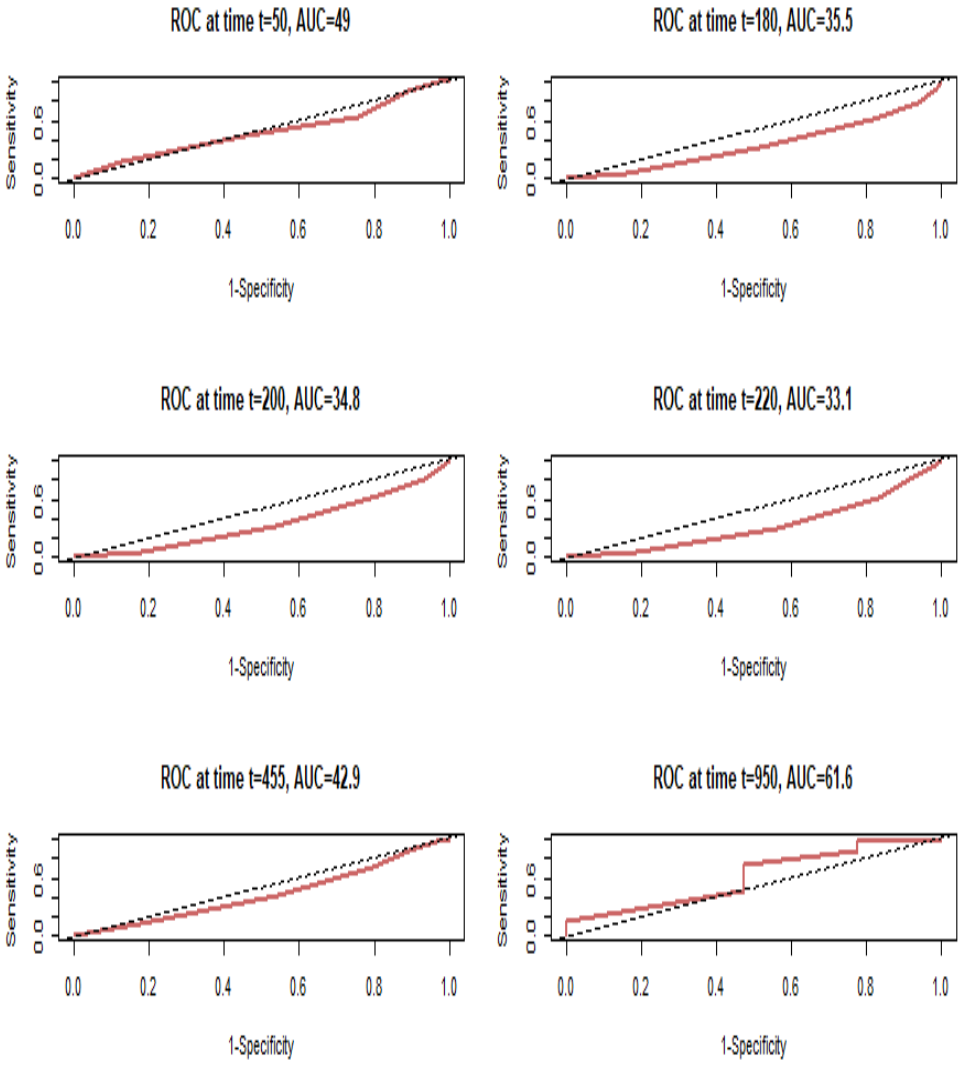
Rysunek 4.54: Time - dependent ROC metodą Coxa dla zmiennej ph.ecog - opracowanie własne



Rysunek 4.55: Time - dependent ROC metodą Coxa dla zmiennej ph.ecog - opracowanie własne

```
ROC.ph.karno <- timeROC(T = lung$time, delta = lung$status,
                        weighting = "cox",
                        marker = lung$ph.karno, cause = 1,
                        times=c(50, 180, 200, 220, 455, 950))
ROC.ph.karno

## Time-dependent-Roc curve estimated using IPCW  (n=226, without competing risks).
##       Cases Survivors Censored AUC (%)
## t=50     11       215        0   48.99
## t=180    61       158        7   35.49
## t=220    79       132       15   33.10
## t=455   131        46       49   42.86
## t=950   163         3       60   61.64
##
## Method used for estimating IPCW:cox
##
## Total computation time : 0.02  secs.
```

Rysunek 4.56: Time - dependent ROC metodą Coxa dla zmiennej ph.karno - opracowanie własne



Rysunek 4.57: Time - dependent ROC metodą Coxa dla zmiennej ph.karno - opracowanie własne

Jak wspomnieliśmy wcześniej AUC, czyli pole pod krzywą ROC jest skalarną reprezentacją oczekiwanej wydajności klasyfikatora. Współczynnik AUC ma wartość z przedziału 0 do 1, przy czym wyższe wartości świadczą o lepszym dopasowaniu klasyfikatora. Mając tę wiedzę, porównywanie krzywych można by ograniczyć jedynie do porównywania wskaźników AUC.

W każdej z poniższych tabel porównaliśmy wartości pola AUC dla różnych zmiennych w różnych przedziałach czasowych. Użyliśmy do tego dwóch estymatorów - Kaplana - Meiera oraz Coxa, w celu omówienia ich zgodności. Pierwszym, ogólnym wnioskiem, jaki możemy wysnuć dla wszystkich tabel, jest fakt, że dla czasu $t = 50$ w każdym przypadku wartość estymatorów Kaplana - Meiera oraz Coxa są takie same.

	Kaplan - Meier	Cox
t = 50	83.69	83.69
t = 180	57.32	57.53
t = 220	52.84	53.35
t = 455	55.17	56.85
t = 950	43.26	48.27

Tabela 4.10: Porównanie wartości pola AUC zależnego od czasu (w %) dla zmiennej wiek - opracowanie własne

Z przeprowadzonej analizy (4.43), (4.51), (4.10) wynika, że w dowolnym momencie czasu t, estymator Coxa przyjmuje większe wartości pola AUC niż estymator Kaplana - Meiera. Dalej widzimy, że najlepsze dopasowanie wydajności klasyfikatora występuje dla zadanego $t = 50$, które stopniowo się zmniejsza dla większej wartości parametru t. Co ważne, dla czasu $t = 950$ otrzymaliśmy wartość pola AUC $< 50\%$, a to oznacza, że nasz klasyfikator jest nieprawidłowy.

	Kaplan - Meier	Cox
t = 50	33.94	33.94
t = 180	38.81	38.57
t = 220	40.47	39.85
t = 455	42.26	39.76
t = 950	50.16	42.05

Tabela 4.11: Porównanie wartości pola AUC zależnego od czasu (w %) dla zmiennej płci - opracowanie własne

Powyższe dane (4.45), (4.53), (4.11) informują, że (odwrotnie niż we wcześniejszym przypadku) estymator Kaplana - Meiera zawsze przyjmuje większe wartości niż estymator Coxa. Należy jednak zauważyć, że dla każdego przypadku wartości pola AUC są małe, zatem oznaczają nieprawidłowe dopasowanie klasyfikatora. W przypadku estymatora Kaplana - Meiera dla czasu $t = 950$ pole AUC $= 50.16$ co dowodzi o losowości klasyfikatora.

Dane zawarte w tabeli (4.12) (oraz na wykresach (4.47), (4.55)) potwierdzają, że wartość estymatora Coxa jest zawsze większa niż wartość estymatora Kaplana - Meiera. Ponadto, w każdym momencie czasowym zauważamy dobre dopasowanie klasyfikatora.

	Kaplan - Meier	Cox
t = 50	55.26	55.26
t = 180	64.68	64.78
t = 220	66.59	66.86
t = 455	56.69	57.55
t = 950	62.43	63.71

Tabela 4.12: Porównanie wartości pola AUC zależnego od czasu (w %) dla zmiennej ph.ecog - opracowanie własne

	Kaplan - Meier	Cox
t = 50	48.99	48.99
t = 180	35.63	35.49
t = 220	33.54	33.10
t = 455	44.95	42.86
t = 950	69.05	61.64

Tabela 4.13: Porównanie wartości pola AUC zależnego od czasu (w %) dla zmiennej ph.karno - opracowanie własne

Na podstawie danych liczbowych zawartych w tabeli (4.13) (oraz na wykresach (4.49), (4.57)) zauważamy, że estymator Kaplana - Meiera przyjmuje większe wartości niż estymator Coxa w każdym momencie czasowym. Najlepsze dopasowanie klasyfikatora występuje dla czasu $t = 950$. Dla pozostałych wartości parametru t, wartość pola AUC < 50, zatem mamy do czynienia ze złym dopasowaniem jakości klasyfikatora.

Rozdział 5

PODSUMOWANIE

W pracy zostały wykonane symulacje badające podobieństwa i różnice w metodach estymacji krzywej ROC w przypadku zależnym od czasu, oraz bez tej zależności. Omówione zostały metody estymacji krzywych ROC w przypadkach, gdy funkcje czułości i swoistości zależą od czasu.

Przeprowadzona została analiza danych z rozkładu log - normalnego, gdzie dla próbki długości $n = 100$ badaliśmy pokrycie się krzywych ROC estymowanych oraz teoretycznych dla różnych parametrów czasu t. Okazało się, że dla zadanej próby oraz $t \leq 1$ krzywe pokrywają się dość dokładnie, zaś dla $t > 1$ zdecydowanie gorzej.

Kolejnym podejmowanym w pracy tematem była analiza danych pochodzących z rozkładu trwałości zmęczeniowej, czyli rozkładu Birnbauma - Saundersa. Pozwoliła nam ona na zbadanie różnych dopasowań krzywych ROC estymowanych i teoretycznych w przypadkach zależnych od parametru kształtu (α) bądź skali (β). Porównane zostało dopasowanie krzywych dla prób różnej długości ($n = 30, 70, 150, 1000$). Zgodnie z przypuszczeniami, większa długość próby wpływa na lepsze dopasowanie krzywych. Wyraźnie widać, że dla próby $n = 70$, estymowana oraz teoretyczna wartość krzywych ROC dla stałej wartości parametru α oraz zmiennego β jest widocznie dokładniejsza, niż w sytuacji odwrotnej.

Ostatnim zagadnieniem poruszanym w pracy była analiza przeżycia na podstawie rzeczywistych danych dostępnych w pakiecie R. Dane te dotyczyły przeżycia pacjentów z zaawansowanym rakiem płuc z North Central Cancer Treatment Group. Przeprowadzona analiza koncentrowała się na oszacowaniu prawdopodobieństwa przeżycia dzięki estymatorom Kaplana - Meiera i Fleminga - Harringtona, testów log - rank oraz nieparametrycznego modelu Coxa. Najistotniejszą częścią pracy była analiza krzywych ROC oraz ROC zależnych od czasu. Dla jego różnych wartości, określonych w dniach, na podstawie wykresów oraz wartości pola AUC szacowaliśmy przeżycie badanej grupy względem różnych charakterystyk zmiennych.

Bibliografia

[1] Akritas, M. G., Bershady, M. A. Linear regression for astronomical data with measurement errors and intrinsic scatter. *arXiv*, 1996.

[2] Bansal, A., Heagerty, P. J. A comparison of landmark methods and time-dependent ROC methods to evaluate the time-varying performance of prognostic markers for survival outcomes. *Diagnostic and Prognostic Research 3*, 1 (July 2019).

[3] Barros, M., Paula, G. A., Leiva, V. An r implementation for generalized birnbaum–saunders distributions. *Computational Statistics &amp Data Analysis 53*, 4 (Feb. 2009), 1511–1528.

[4] Bhattacharyya, G., Fries, A. Fatigue failure models birnbaum-saunders vs. inverse gaussian. *IEEE Transactions on Reliability R-31*, 5 (Dec. 1982), 439–441.

[5] Birnbaum, Z., Saunders, S. Estimation for a family of life distributions with applications to fatigue. *Journal of Applied Probability 6*, 2 (Aug. 1969), 319–327.

[6] Birnbaum, Z., Saunders, S. A new family of life distributions. *Journal of Applied Probability 6*, 2 (Aug. 1969), 319–327.

[7] Breslow, N., Crowley, J. A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics 2*, 3 (May 1974).

[8] Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological) 34*, 2 (Jan. 1972), 187–202.

[9] Das, A. Quick guide to survival analysis using kaplan meier curve, 2020.

[10] Desmond, R. J., Singer, J. L., Singer, D. G., Calam, R., Colimore, K. Family mediation patterns and television viewing. *Human Communication Research 11*, 4 (June 1985), 461–480.

[11] Deszyńska, A. Model hazardów proporcjonalnych coxa, 2011.

[12] Díaz-Coto, S., Martínez-Camblor, P., Pérez-Fernández, S. smoothROCtime: an r package for time-dependent ROC curve estimation. *Computational Statistics 35*, 3 (Jan. 2020), 1231–1251.

[13] Galton, F. *Natural inheritance*. Macmillan, 1889.

[14] Gneiting, T., Walz, E.-M. Receiver operating characteristic (ROC) movies, universal ROC (UROC) curves, and coefficient of predictive ability (CPA). *Machine Learning* (Dec. 2021).

[15] Harańczyk, G. Krzywe roc, czyli ocena jakości klasyfikatora i poszukiwanie optymalnego punktu odcięcia, 2010.

[16] Harańczyk, G. Modelowanie czasu trwania - model proporcjonalnego hazardu coxa, 2011.

[17] Harrell, F. E., Lee, K. L., Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine 15 4* (1996), 361–87.

[18] Harrington, D. P., Fleming, T. R. A class of rank test procedures for censored survival data. *Biometrika 69*, 3 (1982), 553–566.

[19] Hasab, A. A. COVID- 19 screening by RT-PCR: An epidemiological modelling, 2020.

[20] Heagerty, P. J., Lumley, T., Pepe, M. S. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics 56*, 2 (June 2000), 337–344.

[21] Heagerty, P. J., Zheng, Y. Survival model predictive accuracy and ROC curves. *Biometrics 61*, 1 (Mar. 2005), 92–105.

[22] Incerti, D. Parametric survival modeling, 2019.

[23] Kamarudin, A. N., Cox, T., Kolamunnage-Dona, R. Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Medical Research Methodology 17*, 1 (Apr. 2017).

[24] Kang, K., Pan, D., Song, X. A joint model for multivariate longitudinal and survival data to discover the conversion to alzheimer's disease. *Statistics in Medicine 41*, 2 (Nov. 2021), 356–373.

[25] Kaplan, E. L., Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association 53*, 282 (June 1958), 457–481.

[26] Kurniasari, D., Widyarini, R., Warsono, Antonio, Y. Characteristics of hazard rate functions of log-normal distributions. *Journal of Physics: Conference Series 1338*, 1 (Oct. 2019), 012036.

[27] Li, R., Ning, J., Feng, Z. Estimation and inference of predictive discrimination for survival outcome risk prediction models. *Lifetime Data Analysis 28*, 2 (Jan. 2022), 219–240.

[28] Lin, Y., Yang, J. On statistical moments of fatigue crack propagation. *Engineering Fracture Mechanics 18*, 2 (Jan. 1983), 243–256.

[29] Lu, W., Yu, S., Liu, H., Suo, L., Tang, K., Hu, J., Shi, Y., Hu, K. Survival analysis and risk factors in COVID-19 patients. *Disaster Medicine and Public Health Preparedness* (Mar. 2021), 1–6.

[30] Lu, Y.-J., Gong, Y., Li, W.-J., Zhao, C.-Y., Guo, F. The prognostic significance of a novel ferroptosis-related gene model in breast cancer. *Annals of Translational Medicine 10*, 4 (Feb. 2022), 184–184.

[31] Mantel, N. Evaluation of survival data and two new rank order statistics arising in its consideration, 1966.

[32] Narkhede, S. Understanding confusion matrix, 2018.

[33] Noda, Y., Tomita, H., Ishihara, T., Tsuboi, Y., Kawai, N., Kawaguchi, M., Kaga, T., Hyodo, F., Hara, A., Kambadakone, A. R., Matsuo, M. Prediction of overall survival in patients with pancreatic ductal adenocarcinoma: histogram analysis of ADC value and correlation with pathological intratumoral necrosis. *BMC Medical Imaging 22*, 1 (Feb. 2022).

[34] Padgett, W. On bayes estimation of reliability for the birnbaum-saunders fatigue life model. *IEEE Transactions on Reliability R-31*, 5 (Dec. 1982), 436–438.

[35] Pantoja-Galicia, N., Okereke, O. I., Blacker, D., Betensky, R. A. Concordance measures and time-dependent ROC methods. *Biostatistics &amp Epidemiology 5*, 2 (May 2021), 232–249.

[36] Purkayastha, S., Xiao, Y., Jiao, Z., Thepumnoeysuk, R., Halsey, K., Wu, J., Tran, T. M. L., Hsieh, B., Choi, J. W., Wang, D., Vallières, M., Wang, R., Collins, S., Feng, X., Feldman, M., Zhang, P. J., Atalay, M., Sebro, R., Yang, L., Fan, Y., hua Liao, W., Bai, H. X. Machine learning-based prediction of COVID-19 severity and progression to critical illness using CT imaging and clinical data. *Korean Journal of Radiology 22*, 7 (2021), 1213.

[37] Raaijmakers, F. J. M. The lifetime of a standby system of units having the birnbaum and saunders distribution. *Journal of Applied Probability 17*, 2 (June 1980), 490–497.

[38] Salinas-Escudero, G., Carrillo-Vega, M. F., Granados-García, V., Martínez-Valverde, S., Toledano-Toledano, F., Garduño-Espinosa, J. A survival analysis of COVID-19 in the mexican population. *BMC Public Health 20*, 1 (Oct. 2020).

[39] Saunders, S. C. A family of random variables closed under reciprocation. *Journal of the American Statistical Association 69*, 346 (June 1974), 533–539.

[40] Schemper, M., Henderson, R. Predictive accuracy and explained variation in cox regression. *Biometrics 56*, 1 (Mar. 2000), 249–255.

[41] Thomas, B., M, C. V. On a generalized birnbaum saunders distribution, 2021.

[42] Warren F. Kuhfeld, Y. S. Creating and customizing the kaplan-meier survival plot in proc lifetest in the sas, 2014.

[43] Whittaker, E. T., Watson, G. N. *A Course of Modern Analysis*. Cambridge University Press, Sept. 1996.