

**Analiza czasu trwania - Raport Zaliczeniowy**

***Analiza przeżycia pacjentów chorych na raka piersi***

[**Wstęp** 3](#_Toc169559125)

[**Zbiór Danych** 3](#_Toc169559126)

[**Opis** 3](#_Toc169559127)

[**Zmienna cenzurująca** 3](#_Toc169559128)

[**Zmienna objaśniana** 3](#_Toc169559129)

[**Zmienne objaśniające** 4](#_Toc169559130)

[**Kategoryzacja zmiennych ciągłych** 7](#_Toc169559131)

[**Tablice trwania życia** 8](#_Toc169559132)

[**Metoda tradycyjna** 8](#_Toc169559133)

[**Metoda Kaplana-Meiera** 13](#_Toc169559134)

[**Modele Parametryczne** 14](#_Toc169559135)

[**Model wykładniczy** 14](#_Toc169559136)

[**Interpretacja wyników parametrów modelu wykładniczego** 15](#_Toc169559137)

[**Modele Semiparametryczne, podejście klasyczne** 17](#_Toc169559138)

[**Weryfikacja założenia proporcjonalnych hazardów – metoda graficzna** 17](#_Toc169559139)

[**Weryfikacja założenia proporcjonalnych hazardów – reszty Schoenfelda** 18](#_Toc169559140)

[**Weryfikacja założenia proporcjonalnych hazardów – zmienne zależne od czasu** 20](#_Toc169559141)

[**Modele proporcjonalnych i nieproporcjonalnych hazardów** 21](#_Toc169559142)

[**Modele semiparametryczne, podejście bayesowskie** 23](#_Toc169559143)

[**Podsumowanie** 32](#_Toc169559144)

[**Spis tabel i rysunków** 33](#_Toc169559145)

# **Wstęp**

Rak piersi jest jednym z najczęściej diagnozowanych nowotworów u kobiet na całym świecie. Liczne dane epidemiologiczne, zarówno w Polsce jak i za granicą, wskazują na rosnącą zachorowalność w ostatnich latach. Współczesne badania nad przeżyciem chorych na raka piersi koncentrują się na identyfikacji czynników wpływających na wyniki leczenia.

Celem niniejszej pracy jest przeprowadzenie analizy przeżycia wśród pacjentek z rakiem piersi, uwzględniając wybrane cechy zidentyfikowane podczas badania, takie jak rozmiar guza oraz inne charakterystyki pacjentek - między innymi informacje o podjęciu decyzji o terapii hormonalnej.

W poniższym raporcie zostaną przedstawione nieparametryczne, parametryczne oraz semiparametryczne modele przeżycia, których zadaniem będzie zidentyfikowanie głównych czynników wpływających na czas przeżycia pacjentek chorych na raka. Analiza wyników umożliwi podejmowanie skuteczniejszych decyzji dotyczących opieki chorymi.

# **Zbiór Danych**

## **Opis**

Zbiór danych wykorzystany do analizy pochodzi z kolekcji Breast Cancer (METABRIC, Nature 2012 & Nat Commun 2016), dostępnej na platformie cBioPortal. Jest to kompleksowy zbiór danych pochodzący z badania wykonanego dla czasopisma Nature obejmujący informacje o uczestników, u których zdiagnozowano raka piersi. Zawiera on szczegółowe dane dotyczące 2509 pacjentów, opisanych za pomocą 22 zmiennych, co stanowi solidną podstawę do przeprowadzenia szeroko zakrojonej analizy przeżycia i badań nad rokowaniem w raku piersi.

## **Zmienna cenzurująca**

W analizowanym zbiorze danych, zmienna cenzurująca została określona jako “V”. Wskazuje ona, czy osoba objęta badaniem dożyła do jego zakończenia i zmarła z przyczyn niezwiązanych z badanym zjawiskiem, czy też zmarła w wyniku tego zjawiska. Zmienna ta dostarcza informacji o długości czasu obserwacji dla każdego uczestnika badania, co jest kluczowe dla precyzyjnej analizy czasu do wystąpienia interesującego nas zdarzenia. Umożliwia to uwzględnienie w analizie takich czynników jak śmierć z innych przyczyn, które mogą mieć wpływ na wyniki oraz ich interpretację.

|  |  |  |
| --- | --- | --- |
| **Zmienna** | **Wartość** | **Opis** |
| *V* | 0 | Osoby które przeżyły badanie i które zmarły z innych powodów |
| 1 | Osoby które zmarły |

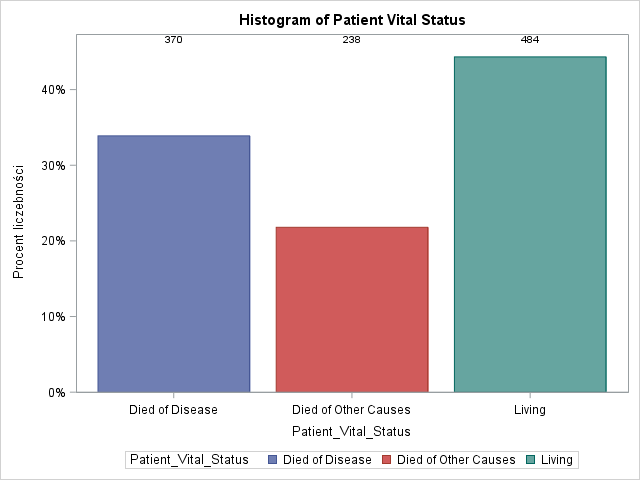
Tabela 1: Szczegóły zmiennej cenzurującej

## **Zmienna objaśniana**

W badanym zbiorze danych, zmienną objaśnianą jest zmienna ‘Patient\_Vital\_Status’ klasyfikuje uczestników na trzy kategorie: osoby, które zmarły na skutek choroby, osoby, które zmarły z innych przyczyn oraz osoby, które przeżyły do końca okresu badawczego. Ta zmienna jest głównym wskaźnikiem dla oceny wyników badań nad przeżyciem i dla analizy wpływu czynników prognostycznych na długość życia pacjentów. Dokładne ustalenie statusu życiowego każdego pacjenta jest niezbędne dla formułowania wniosków z przeprowadzonego badania, umożliwiając szczegółowe rozpatrzenie skuteczności terapii i jakości opieki zdrowotnej w kontekście choroby.

|  |  |  |
| --- | --- | --- |
| **Zmienna** | **Wartość** | **Opis** |
| *Patient\_Vital\_Status* | Died of Disease | Osoby, które zmarły z powodu choroby |
| Died of Other Causes | Osoby, które zmarły z innych powodów |
| Living | Osoby żyjące pod koniec okresu badania |

Tabela 2: Szczegóły zmiennej objaśnianej



Rysunek 1: Histogram zmiennej objaśnianej

Rozkład pacjentów w poszczególnych kategoriach świadczy o zbalansowanej próbie, co jest pożądane w badaniach statystycznych i przyczynia się do wiarygodności wyników. Wysoki odsetek osób żyjących na koniec okresu badania sugeruje skuteczność stosowanych terapii lub wskazuje na potrzebę dalszych badań w celu zrozumienia czynników wpływających na przeżycie. Z kolei obserwacja, że znacząca liczba pacjentów zmarła z innych przyczyn niż choroba, podkreśla znaczenie uwzględniania różnorodnych czynników prognostycznych w analizie przeżycia.

## **Zmienne objaśniające**

|  |  |  |
| --- | --- | --- |
| **Zmienna** | **Wartość** | **Opis** |
| *Type\_of\_Breast\_Surgery* | Breast Conserving | Operacja oszczędzająca pierś |
| Mastectomy | Całkowite usunięcie piersi |
| *Cancer\_Type\_Detailed* | Breast | Rak piersi, generalna kategoria obejmująca wszystkie typy nowotworów zlokalizowanych w tkance piersiowej |
| Breast Invasive Ductal Carcinoma | Najczęstszy typ raka piersi, który rozwija się w kanalach mlekowych piersi, a następnie może przenikać do otaczającej tkanki |
| Breast Invasive Lobular Carcinoma | Nowotwór rozwijający się w komórkach zrazikowych piersi, której cechą jest inwazja i potencjalne rozprzestrzenianie się do innych obszarów ciała |
| Breast Invasive Mixed Mucinous Carcinoma | Rzadki typ raka piersi, w którym komórki rakowe produkują śluz; może zawierać elementy innych typów raka piersi |
| Breast Mixed Ductal and Lobular Carcinoma | Nowotwór zawierający zarówno komórki rakowe pochodzące z przewodów mlekowych, jak i zrazików piersiowych, często w różnych proporcjach |
| *Cellularity* | High | Duża ilość komórek w tkance |
| Low | Mała ilość komórek w tkance |
| Moderate | Umiarkowana ilość komórek w tkance |
| *Chemotherapy* | No | Osoby, które nie miały chemioterapii |
| Yes | Osoby, które przeszły chemioterapię |
| *Pam50\_and\_Claudin\_low\_subtype* | Basal | Potrójnie ujemny (ER-, PR-, HER2-), agresywny, częsty u młodszych kobiet |
| Her2 | Wysoka ekspresja HER2, agresywny, odpowiada na terapie celowane HER2 |
| LumA | Wysoka ekspresja ER i PR, niski wskaźnik proliferacji, dobra odpowiedź na hormonoterapię, najlepsze rokowania |
| LumB | Wysoka ekspresja ER i PR, wyższa ekspresja HER2 niż Luminal A, wyższy wskaźnik proliferacji, bardziej agresywny |
| NC | Niesklasyfikowany, nie pasuje jednoznacznie do żadnego z innych podtypów |
| Normal | Podobny do Luminal A, ale mniej zrozumiały, dobre rokowania |
| Claudin-low | Niska ekspresja genów adhezji komórek, cechy komórek macierzystych nowotworowych, agresywny |
| *Cohort* | 1 | SEER |
| 2 | METABRIC |
| 3 | TCGA |
| 5 | FUSCC |
| *ER\_status\_measured\_by\_IHC* | Negative | Brak receptorów estrogenowych |
| Positive | Występują receptory estrogenowe |
| *Neoplasm\_Histologic\_Grade* | 1 | Najmniejszy stopień agresywności. |
| 2 | Umiarkowany stopień agresywności |
| 3 | Najwyższy stopień agresywności |
| *HER2\_status\_measured\_by\_SNP* | Gain | W komórkach nowotworowych występuje nadmierna liczba kopii genu HER2 |
| Loss | W komórkach nowotworowych brakuje części lub całości genu HER2 |
| Neutral | W komórkach nowotworowych występuje standardowa liczba kopii genu HER2 |
| Undef | Oznacza, że status genu HER2 nie został określony lub wyniki są niejednoznaczne |
| *HER2\_Status* | Negative | Komórki nowotworowe nie wykazują nadekspresji białka HER2 |
| Positive | Komórki nowotworowe wykazują nadekspresję białka HER2 |
| *Hormone\_Therapy* | 0 | Osoby, które nie przeszły terapii hormonalnej |
| 1 | Osoby, które przeszły terapii hormonalnej |
| *Inferred\_Menopausal\_State* | Post | Osoby po menopauzie |
| Pre | Osoby przed menopauzą |
| *Integrative\_Cluster* | 1-12 | Klasyfikacja nowotworu na podstawie ekspresji genów |
| *Primary\_Tumor\_Laterality* | Left | Rak w lewej piersi |
| Right | Rak w prawej piersi |
| *Oncotree\_Code* | BREAST | Ogólne określenie odnoszące się do nowotworów piersi |
| IDC | Rak przewodowy inwazyjny |
| ILC | Rak zrazikowy inwazyjny |
| IMMC | Inwazyjny rak mikropapilarny |
| MDLC | Rak rdzeniasty |
| *PR\_Status* | Negative | Brak receptorów progesteronowych na powierzchni komórek nowotworowych |
| Positive | Obecność receptorów progesteronowych na powierzchni komórek nowotworowych |
| *Radio\_Therapy* | 0 | Osoby, które nie przeszły radioterapii. |
| 1 | Osoby, które przeszły radioterapię |
| *3-gene\_classifier\_subtype* | ER+/HER2- High Prolif | Podtyp nowotworu, który jest dodatni pod względem receptorów estrogenowych (ER+), ale ma wysoki wskaźnik proliferacji komórkowej |
| ER+/HER2- Low Prolif | Podtyp nowotworu ER+, ale z niską proliferacją komórek |
| ER-/HER2- | Nowotwór jest ujemny pod względem receptorów estrogenowych (ER-) i HER2 (HER2-) |
| HER2+ | Nowotwór jest dodatni pod względem HER2 |
| *Tumor\_stage* | 1 | Nowotwór jest mały i ograniczony do miejsca, gdzie się rozpoczął |
| 2 | Nowotwór jest większy niż w stadium 1, ale nadal może być ograniczony do miejsca początkowego lub rozprzestrzenił się na pobliskie tkanki |
| 3 | Nowotwór jest bardziej zaawansowany i rozprzestrzenił się głębiej w pobliskie tkanki |
| 4 | Nowotwór jest najbardziej zaawansowany i rozprzestrzenił się na odległe części ciała (przerzuty) |
| *Tumor\_Other\_Histologic\_Subtyp* | Ductal/NST | Rak przewodowy lub typ raka piersi, który nie wykazuje cech charakterystycznych dla innych podtypów |
|  | Lobular | Rak zrazikowy |
| Mixed | Nowotwór mieszany, który zawiera elementy różnych typów histologicznych |
| Other | Inne rzadkie typy raka piersi |

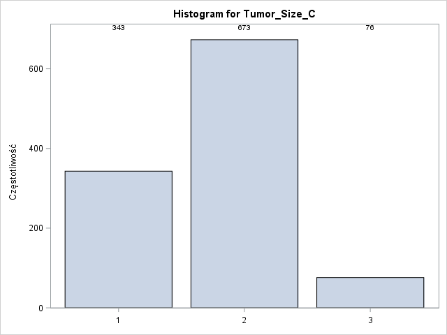
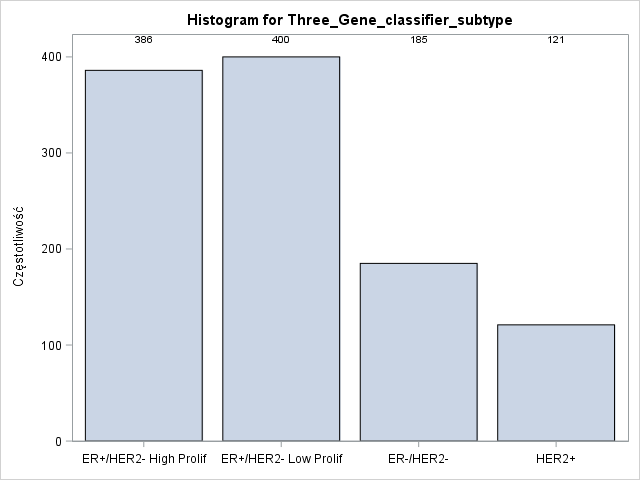
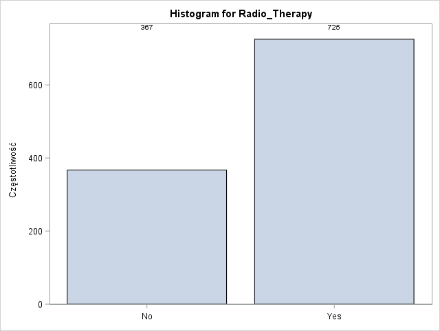
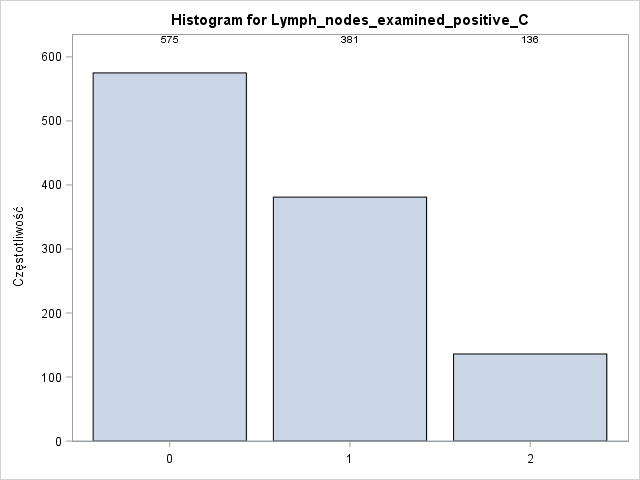
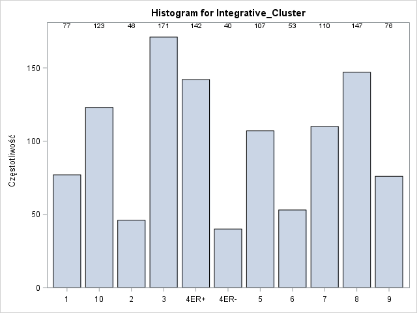
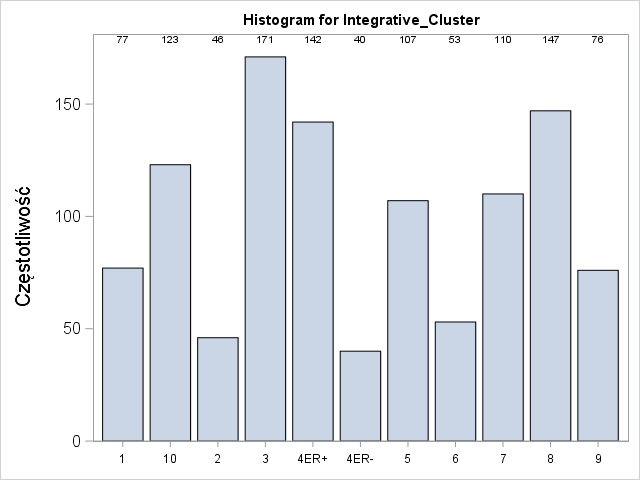
Tabela 3: Szczegóły zmiennych objaśniających

## **Kategoryzacja zmiennych ciągłych**

Kategoryzacja zmiennych to proces dzielenia danych na klarowne i użyteczne kategorie w celu ułatwienia analizy i interpretacji. Kategoryzacja zastosowana w badaniu oparta jest na systemie TNM, który jest powszechnie stosowany w onkologii do klasyfikacji nowotworów na podstawie trzech głównych czynników: rozmiaru guza (T), zaawansowania w węzłach chłonnych (N) oraz obecności przerzutów odległych (M). Ten system umożliwia dokładne określenie stadium choroby, co jest kluczowe dla odpowiedniego planowania leczenia, prognozowania wyników oraz porównywania danych klinicznych między pacjentami. Dzięki kategoryzacji za pomocą systemu TNM możliwe jest lepsze zrozumienie dynamiki choroby nowotworowej oraz skuteczniejsze podejmowanie decyzji terapeutycznych.

|  |  |  |
| --- | --- | --- |
| **Zmienna** | **Wartość** | **Opis** |
| *Lymph\_nodes\_examined\_positive\_C* | 0 | 0 węzłów chłonnych, w których stwierdzono obecność komórek nowotworowych |
| 1 | 1-5 węzłów chłonnych, w których stwierdzono obecność komórek nowotworowych |
| 2 | Więcej niż 5 węzłów chłonnych, w których stwierdzono obecność komórek nowotworowych |
| *Mutation\_Count\_C* | 1 | Stwierdzono istotne mutacje w mniej niż pięciu różnych genach |
| 2 | Stwierdzono istotne mutacje w pięciu lub więcej różnych genach i mniej niż dziesięciu |
| 3 | Stwierdzono istotne mutacje w dziesięciu lub więcej różnych genach |
| *Age\_at\_Diagnosis\_C* | 2 | Osoby poniżej 50 roku życia |
| 3 | Osoby między 50 i 60 rokiem życia |
| 4 | Osoby między 60 i 70 rokiem życia |
| 5 | Osoby powyżej 70 roku życia |
| *Tumor\_Size\_C* | 1 | Guz mniejszy niż 2 cm |
| 2 | Guz pomiędzy 2 cm a 5 cm |
| 3 | Guz większy niż 5 cm |
| *Nottingham\_prognostic\_index\_C* | 1 | NPI < 3.5 - dobra prognoza |
| 2 | NPI >= 3.5 i NPI < 4.5 - średnia prognoza |
| 3 | NPI >= 4.5 i NPI < 5.5 - niska prognoza |
| 4 | NPI >= 5.5 - bardzo niska prognoza |

Tabela 4: Kategoryzacja zmiennych objaśniających



Rysunek 2: Histogramy zmiennych objaśniających

# **Tablice trwania życia**

Tablice trwania życia, znane również jako tablice przeżycia, są fundamentalnym narzędziem w analizie czasu trwania. Przedstawiają one prawdopodobieństwo przeżycia osób z różnymi cechami, od momentu rozpoczęcia obserwacji do wystąpienia zdarzenia, którym w badanym obszarze jest śmierć. Tablice umożliwiają graficzne i liczbowe przedstawienie czasu, jaki upłynął do zdarzenia, oraz pozwalają na porównanie przeżycia między różnymi grupami pacjentów.

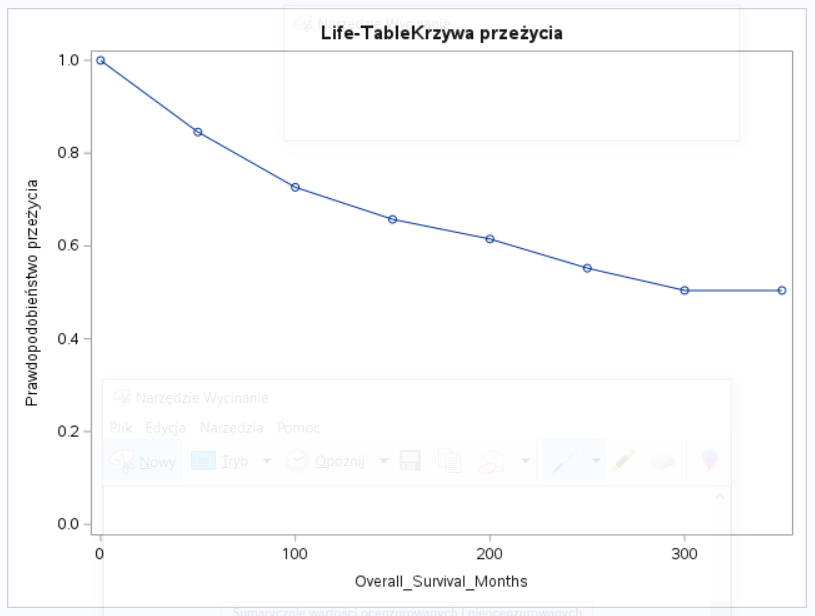
Na podstawie badanego zbioru danych stworzono tablice trwania życia metodą tradycyjną oraz Kaplana-Meiera.

## **Metoda tradycyjna**



Tabela 5: Wartości tabeli przeżycia opracowane metodą tradycyjną

W ramach pierwszych 50 miesięcy obserwacji, zgon zarejestrowano u 164 uczestników badania. W dalszych przedziałach czasowych liczba ta sukcesywnie maleje, osiągając wartość zero w przedziale od 300 do 350 miesięcy. Warunkowe prawdopodobieństwo wystąpienia zdarzenia niepożądanego jest najwyższe dla początkowego przedziału czasowego i wynosi 0,1544, przy standardowym błędzie wynoszącym 0,0111. Wartość tego wskaźnika obniża się w kolejnych przedziałach czasowych, by ponownie wzrosnąć w okresach od 200 do 250 miesięcy oraz od 250 do 300 miesięcy. Prawdopodobieństwo, że pacjent nie doświadczy zgonu z powodu raka piersi w ciągu 300 miesięcy lub dłużej przekracza 50 procent, przy standardowym błędzie oszacowanym na poziomie 0,03.

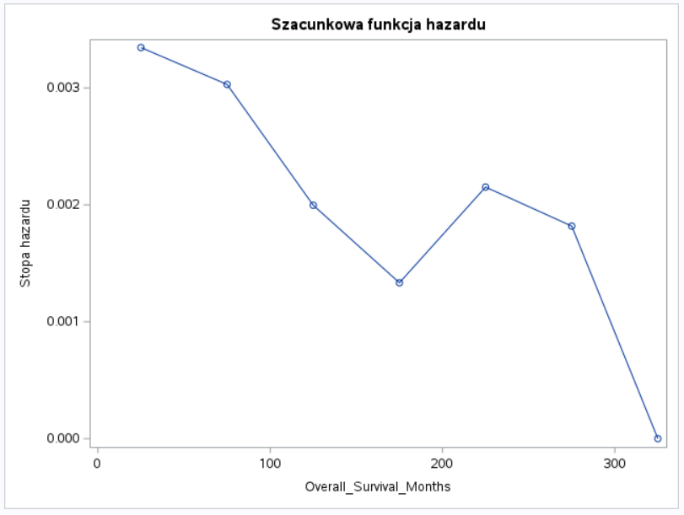


Rysunek 3: Krzywa przeżycia dla zmiennej Overall\_Survival\_Months

Funkcja gęstości prawdopodobieństwa osiąga swoje maksimum w pierwszym przedziale czasowym, z wartością 0,00309 oraz błędem standardowym wynoszącym 0,000222. Funkcja ryzyka dla tego samego przedziału została wyestymowana na poziomie 0,003345, z błędem standardowym równym 0,00026. Wartości obu funkcji wykazują tendencję spadkową w następnych trzech przedziałach czasowych, jednakże obserwuje się wzrost ich wartości w przedziale od 200 do 250 miesięcy. Analogicznie, błędy standardowe tych funkcji również wzrastają. Powyższe tendencje są odzwierciedlone na odpowiadających im wykresach.



Rysunek 4: Funkcja gęstości prawdopodobieństwa dla zmiennej Overall\_Survival\_Months



Rysunek 5: Funkcja hazardu dla zmiennej Overall\_Survival\_Months

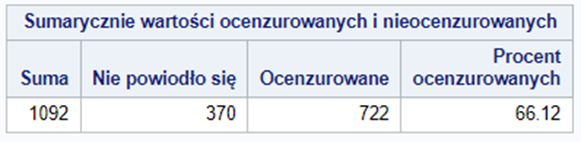
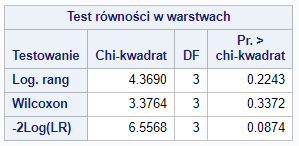


Tabela 6: Podsumowanie wartości ocenzurowanych i nieocenzurowanych

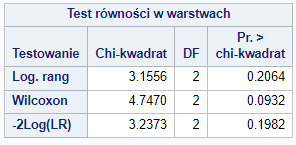
W analizowanym zbiorze danych, zgon spowodowany rakiem piersi zaobserwowano w 370 przypadkach, podczas gdy status ocenzurowany przypisano 722 osobom, co stanowi 66,12% badanej populacji.

Przeprowadzono również analizę wpływu niektórych zmiennych takich jak Age\_at\_diagnosis\_c, Cancer\_Type\_Detailed, Cellularity, Three\_Gene\_classifier\_subtype, Tumor\_size\_C, Chemotherapy, Radio\_Therapy oraz Hormone\_Therapy na krzywe przeżycia. Testy log-rank, Wilcoxona oraz -2Log(LR) nie wykazały statystycznie istotnego wpływu wymienionych zmiennych opisujących leczenie na funkcje przeżycia, z wyjątkiem zmiennej Chemotherapy. Analiza wykazała, że chemioterapia była czynnikiem zwiększającym ryzyko zgonu z powodu raka piersi. Może to sugerować, że chemioterapia jest stosowana głównie w bardziej zaawansowanych stadiach choroby, które wiążą się z wyższym ryzykiem śmiertelności.



*Tabela 7: Testy dla zmiennych Age\_at\_Diagnosis\_c, Hormone\_Therapy, Radio\_Therapy*

Zmienne Cancer\_Type\_Detailed oraz Cellularity wykazały niższe wartości p-value w testach równości rozkładów w warstwach, jednakże nie osiągnęły one poziomu istotności statystycznej na poziomie 0,05. W związku z tym, nie można stwierdzić istotnych statystycznie różnic między funkcjami przeżycia dla poszczególnych grup badanych.

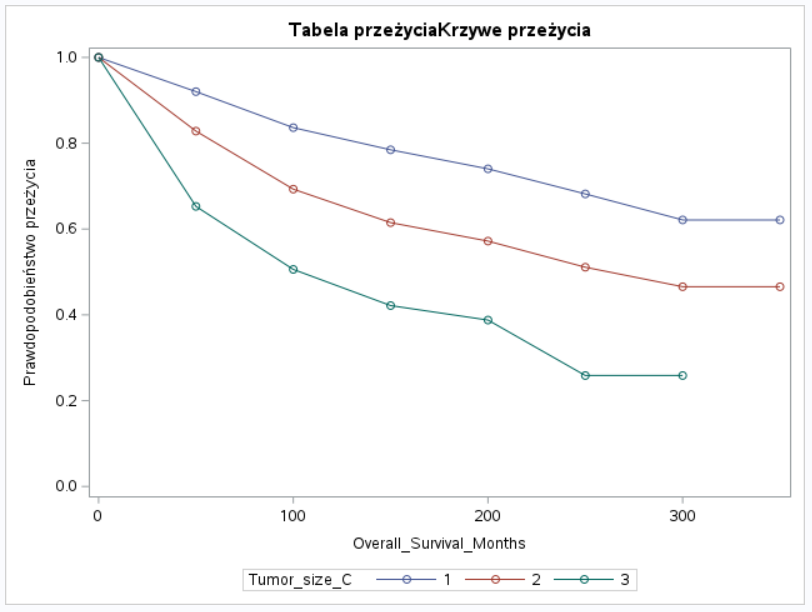


*Tabela 8: Testy dla zmiennych Chemotherapy, Cancer\_Type\_Detailed, Cellularity*

Zmienne Tumor\_size\_C i Three\_Gene\_classifier\_subtype wykazały istotną statystycznie różnicę między krzywymi przeżycia i funkcjami hazardu.

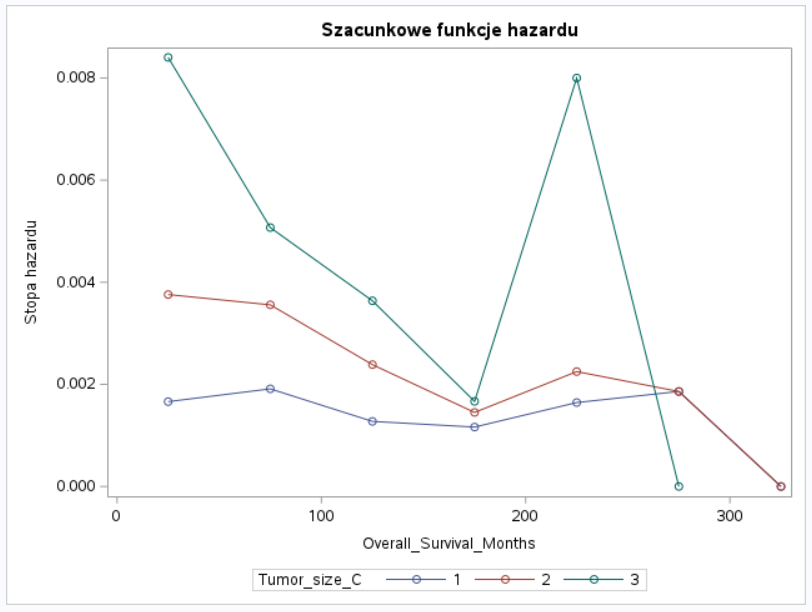


*Tabela 9: Testy dla zmiennych Tumor\_size\_c i Three\_Gene\_classifier\_subtype*



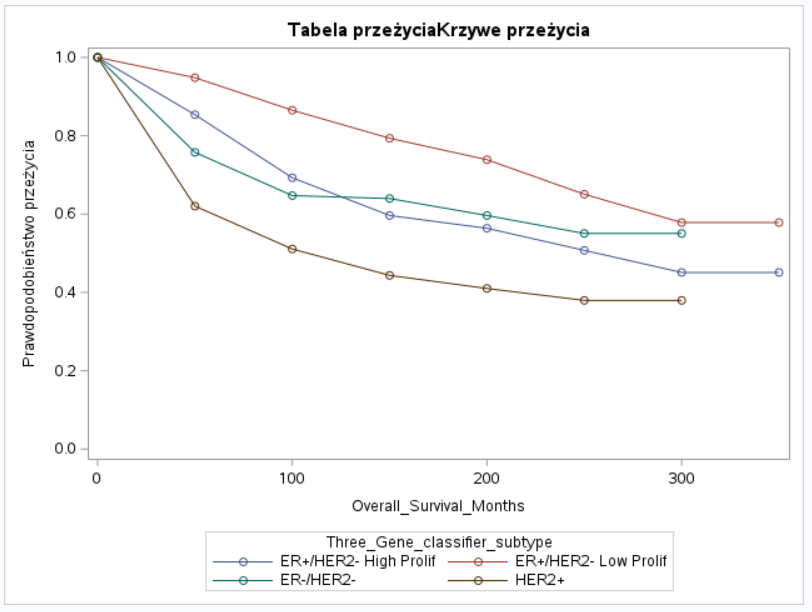
Rysunek 6: Krzywa przeżycia dla zmiennej Tumor\_size\_C

Zmienna Tumor\_size\_C ma bardzo duży wpływ na krzywą przeżycia. Prawdopodobieństwo przeżycia jest największe dla guza małych rozmiarów w przypadku wszystkich przedziałów czasowych.



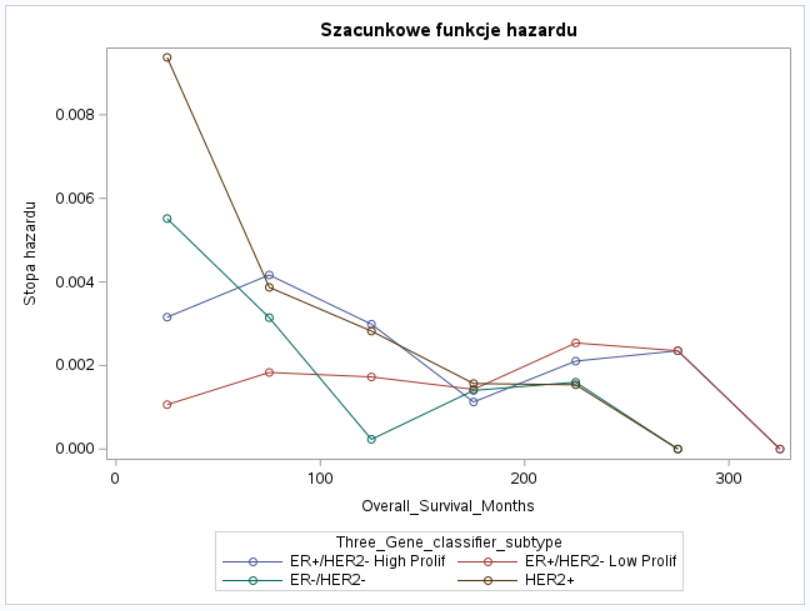
Rysunek 7: Funkcja hazardu dla zmiennej Tumor\_size\_C

Różni się także kształt funkcji hazardu. Dla przedziałów czasowych od 0 do 250 miesięcy funkcja hazardu jest największa dla największej wartości zmiennej tumor\_size\_c. Zmienia się to dla przedziałów powyżej 250 miesięcy, co jednak może być spowodowane małą liczebnością próby. Widać także, że wzrost funkcji hazardu dla całej próby w przedziale 200-250 miesięcy mógł być spowodowany gwałtownym wzrostem funkcji hazardu dla grupy charakteryzującej się zmienną tumor\_size\_c równą 3.



Rysunek 8: Funkcja przeżycia dla zmiennej Three\_Gene\_classifier\_subtype

Zmienna Three\_Gene\_classifier\_subtype także ma duży wpływ na funkcję przeżycia. Dla badanych posiadającyh gen ER+/HER2- Low Prolif prawdopodobieństwo przeżycia jest największe dla wszystkich przedziałów czasowych, a osoby z genem HER2+ mają najmniejsze prawdopodobieństwo przeżycia w tych samych okresach. Z kolei funkcja przeżycia osób z genem ER-/HER2- charakteryzuje się dużym spadkiem w początkowych przedziałach czasowych, przez co prawdopodobieństwo przeżycia jest niższe niż osób z genem ER+/HER2- High Prolif, ale następnie wzrasta po 100. miesiącu powyżej poziomu osób z tym genem.



Rysunek 9: Funkcja hazardu dla zmiennej Three\_Gene\_classifier\_subtype

Kształt funkcji hazardu dla osób z różnymi genami bardzo się różnił. Funkcje hazardu dla grupy z genem HER2+ charakteryzuje bardzo duża wartość początkowa, a następnie duży spadek, nie ma też wzrostu hazardu dla późniejszych przedziałów czasowych. Grupa z genem ER+/HER2- Low Profit ma najmniejszą wartość początkową funkcji hazardu, następnie funkcja się zwiększa, a wartość maksymalną uzyskuje dla przedziału czasowego 200-250 miesięcy. Funkcja hazardu badanych z genem ER-/HER2- spada bardzo szybko w pierwszych przedziałach czasowych, następnie rośnie od przedziału 100-150 miesięcy. Dla grupy z genem ER+/HER2- High Profit funkcja na początku rośnie, następnie maleje, i zaczyna rosnąć po przedziale 150-200 miesięcy.

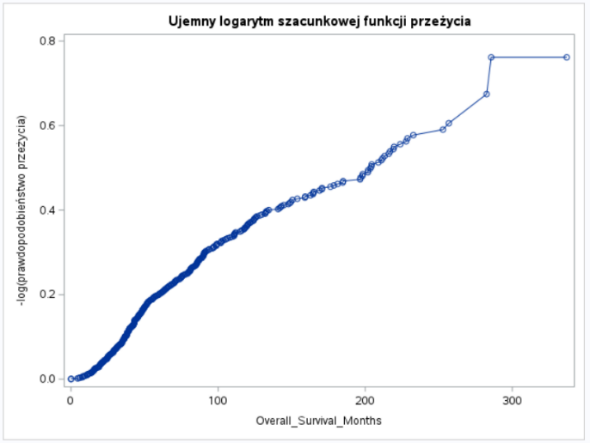
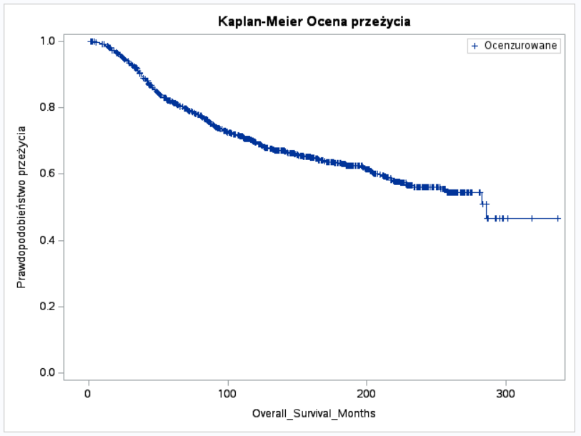
## **Metoda Kaplana-Meiera**

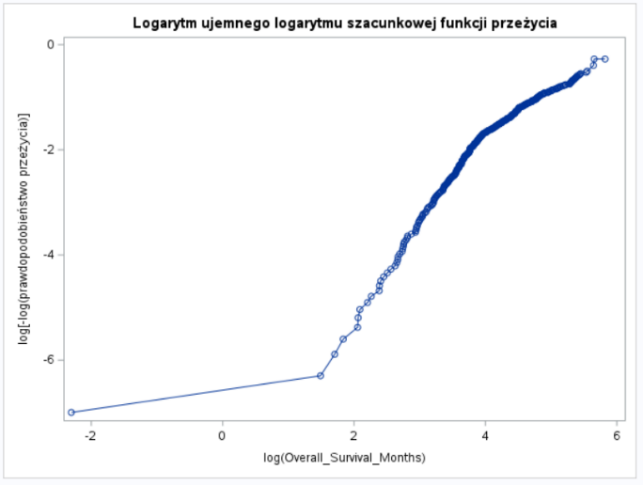
Opracowano także tablice trwania życia metodą Kaplana-Meiera i otrzymano następujące estymowane wartości:



Tabela 10: Wartości tabeli przeżycia opracowane metodą Kaplana-Meiera

Średnie dalsze trwanie życia zostało oszacowane na 200,917 miesiąca, a błąd standardowy wyniósł 3,505. Pierwszy kwartyl wyniósł 89,100, a jego dolna i górna granica kolejno 79,867 i 102,300. Mediana przeżycia wyniosła 285,700 miesiąca.





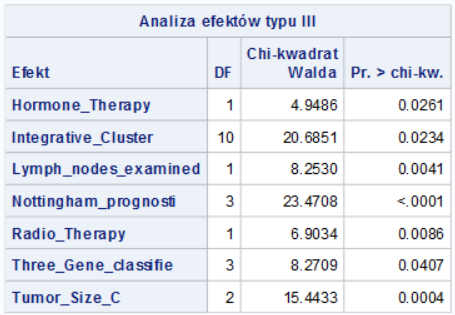
Rysunek 10: Funkcje dla metody Kaplana-Meiera

# **Modele Parametryczne**

W niniejszej części raportu oszacowano parametry modelu parametrycznego, opierając się na klasycznym podejściu. Podejście to opiera się na częstościowej definicji prawdopodobieństwa, zgodnie z którą prawdopodobieństwo danego zdarzenia losowego jest przybliżane przez częstość jego występowania w dużej liczbie identycznych prób, a parametry modelu są nieznanymi, ale określonymi stałymi. Zgodnie z tą definicją, estymatory parametrów modeli uzyskane za pomocą klasycznych metod bazują wyłącznie na informacjach zawartych w próbie losowej.

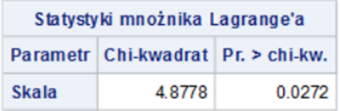
## **Model wykładniczy**

Modelem, na podstawie którego dokonano estymacji parametrów w naszym raporcie, jest model wykładniczy. Poniżej jest przedstawiona tabela analizy efektów typu III dla tego modelu. Przy poziomie istotności 0.05 mamy podstawy do odrzucenia hipotezy zerowej w teście Chi-kwadrat Walda w przypadku wszystkich zmiennych, czyli przyjmujemy, że wszystkie wybrane zmienne objaśniające w naszym modelu są istotne statystycznie dla estymowanych parametrów modelu wykładniczego.



*Tabela 11: Analiza efektów typu III dla modelu wykładniczego*

Podsumowując, nie musimy usuwać żadnych zmiennych, ponieważ na podstawie wyników analizy efektów typu III estymacji parametrów modelu wykładniczego wszystkie zmienne objaśniające okazały się istotne.

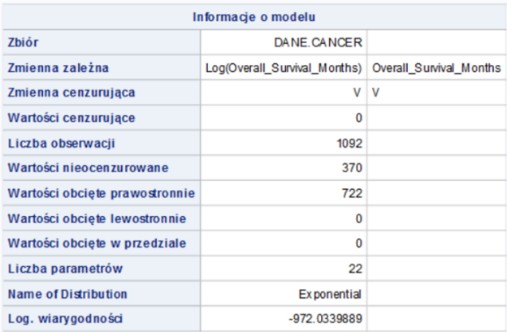


*Tabela 12: Test mnożnika Lagrange'a dla modelu wykładniczego*

Korzystając ze statystyki mnożnika Lagrange’a, możemy sprawdzić założenie o stałości funkcji hazardu. Hipoteza zerowa (H0) zakłada, że funkcja hazardu jest stała w czasie, podczas gdy hipoteza alternatywna (H1) sugeruje, że funkcja hazardu zmienia się w czasie. W naszym przypadku statystyka mnożnika Lagrange’a wynosi 4,8778. Przy poziomie istotności 𝛼 = 0,05 musimy odrzucić hipotezę zerową o stałości funkcji hazardu i przyjąć H1, co oznacza że model wykładniczy nie jest odpowiedni do opisu badanego procesu.

## **Interpretacja wyników parametrów modelu wykładniczego**

Mimo że model wykładniczy nie jest odpowiedni do opisu badanego procesu w przypadku naszych danych, zdecydowano się przedstawić interpretację uzyskanych parametrów, aby pokazać, jak poszczególne zmienne wpływają na wyniki modelu. Należy jednak uwzględnić potencjalne błędy interpretacyjne wynikające z niedopasowania modelu do danych. W poniższej tabeli są przedstawione ogólne informacje o modelu wykładniczym.

  
*Tabela 13: Informacje o modelu wykładniczym*

W tym modelu zmienną zależną jest Overall\_Survival\_Months oraz jej logarytmiczna transformacja Log(Overall\_Survival\_Months). Liczba obserwacji wynosi 1092, z czego 370 przypadków nie jest ocenzurowanych, a 722 przypadki są ocenzurowane prawostronnie. Nie występują wartości ocenzurowane lewostronnie ani w przedziale. Model wykorzystuje 22 parametry. Wybrana dystrybucja to wykładnicza, co oznacza, że model zakłada stałą funkcję hazardu. Logarytm wiarygodności modelu wynosi -972.033989, co jest miarą dopasowania modelu do danych. Niska wartość logarytmu wiarygodności sugeruje, że model może nie być idealnie dopasowany do danych. Warto również zwrócić uwagę na potencjalne błędy interpretacyjne wynikające z użycia modelu wykładniczego, ponieważ założenie o stałej funkcji hazardu nie jest spełnione.

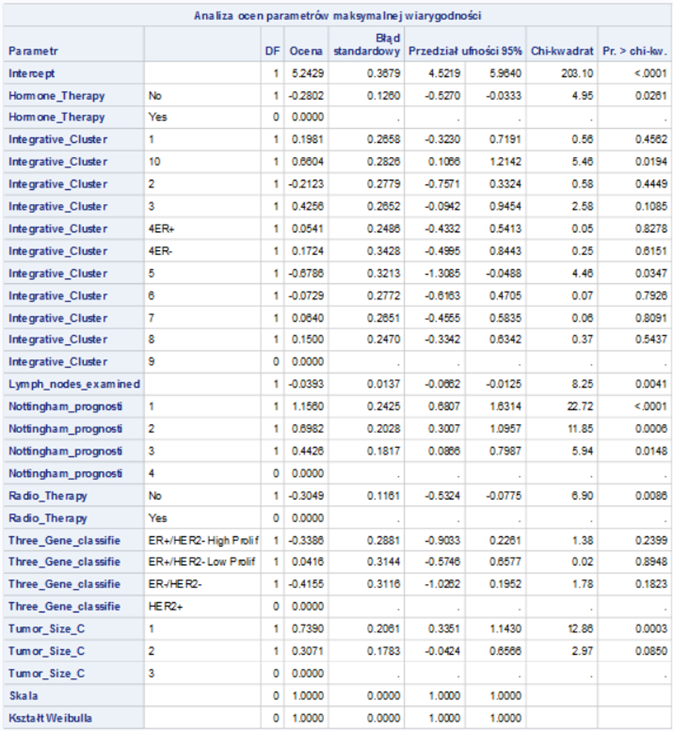
Poniższe tabele przedstawiają statystyki dopasowania modelu wykładniczego w analizie przeżycia. Pierwsza tabela zawiera statystyki dopasowania modelu wykładniczego dla zmiennej zależnej Overall\_Survival\_Months. Druga tabela pokazuje statystyki dopasowania modelu wykładniczego dla odpowiedzi nierejestrowanej.



*Tabela 14: Statystyki dopasowania modelu wykładniczego dla zmiennej zależnej* *Overall\_Survival\_Months*

  
*Tabela 15: Statystyki dopasowania modelu wykładniczego dla odpowiedzi nierejestrowanej*

Poniższa tabela zawiera oceny parametrów maksymalnej wiarygodności dla modelu wykładniczego.



*Tabela 16: Analiza ocen parametrów maksymalnej wiarygodności*

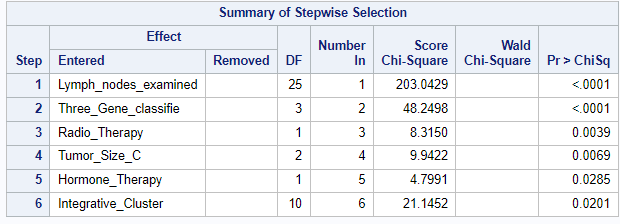
Do oszacowania zostały wybrane niektóre zmienne z modelu. Zmienna binarna Hormone\_Therapy pokazuje, czy dana kobieta przyjmowała lub przymuje terapię hormonalną („Yes” – tak, „No” – nie) w związku z zachorowaniem na raka piersi. W przypadku wartości „No” exp(-0.2802) = 0.756, to oznacza, że średni czas przeżycia kobiet, które nie przyjmowały terapii hormonalnej, był o około 24.4% krótszy niż dla tych kobiet, które przyjmowały terapię hormonową. Co więcej, ryzyko śmierci dla kobiet niekorzystających z terapii hormonalnej było o około (exp(0,28) - 1) ∙ 100% = 32.3% większe niż dla tych, które z terapii korzystały. Zmienna Nottingham\_prognostic jest zmieną kategoryczną, która reprezentuje wynik testu Nottingham Prognostic Index (NPI). NPI jest systemem punktacyjnym, który służy do przewidywania ogólnego przeżycia pacjentek z danym rodzajem raka piersi. Pogrupowaliśmy tą zmienną na 4 kategorie, oznaczone od 1 do 4. Jeśli zmienna przyjmuje wartość 1, to exp(1.1560) = 3.1772, co oznacza, że średni czas przeżycia kobiet, należących do grupy 1 zmiennej Nottingham\_prognostic był o około 3 razy większy niż dla tych kobiet, które należą do grupy 4. Zmienna Three\_Gene\_Classifier jest zmienną kategoryczną, która reprezentuje wynik testu 3-genowego klasyfikatora. 3-genowy klasyfikator jest testem genetycznym, który służy do przewidywania rokowania pacjentek z rakiem piersi. Istnieja 4 wyniki tego testu:

* ER+/HER2-High Prolif 1,
* ER+/HER2-Low Prolif 1,
* ER/HER2-,
* HER2+.

W przypadku wartości „ER+/HER2-High Prolif 1”, exp(-0.3386) = 0.713, co oznacza, że średni czas przeżycia kobiet, których wynik testu genetycznego wskazał wartość „ER+/HER2-High Prolif 1” był o około 28.7% krótszy niż dla tych kobiet, u których wynik badania genetycznego wskazał wartość “HER2+”.

# **Modele Semiparametryczne, podejście klasyczne**

Nasza analiza opiera się na modelach semiparametrycznych, zarówno w ujęciu bayesowskim, jak i klasycznym. Okazują się one niezwykle przydatne, gdy opis zależności hazardu od czasu jest skomplikowany i niemożliwy do analitycznego przedstawienia. Pierwszy krok w budowie modeli semiparametrycznych z wykorzystaniem podejścia klasycznego to dobór zmiennych do modelu. W naszym przypadku dobór zmiennych został dokonany za pomocą metody Stepwise Selection.

  
*Tabela 17: Analiza ocen parametrów maksymalnej wiarygodności*

W dalszej analizie modeli klasycznych w ujęciu semiparametrycznym zostaną wykorzystane tylko zmienne Hormone\_Therapy Lymph\_nodes\_examined\_positive Integrative\_Cluster Radio\_Therapy Three\_Gene\_classifier\_subtype Tumor\_Size\_C, ponieważ te zmienne zostały wybrane przez metodę Stepwize Selection jako najistotniejsze w naszych danych.

## **Weryfikacja założenia proporcjonalnych hazardów – metoda graficzna**

Założenie proporcjonalnych hazardów jest fundamentalne w analizie czasu do zdarzenia w modelach regresji Coxa. Oznacza ono, że wpływ zmiennych objaśniających na hazard jest stały w czasie. Innymi słowy, stosunek hazardów dla różnych wartości danej zmiennej objaśniającej pozostaje taki sam w każdym momencie obserwacji. Aby sprawdzić, czy założenie proporcjonalnych hazardów (PH) jest spełnione, wykorzystano metodę graficzną z wykresem ujemnego logarytmu z ujemnego logarytmu oszacowanej funkcji przeżycia. Na wykresie tym bada się równoległość linii dla zmiennych jakościowych.

Na wykresie ujemnego logarytmu z ujemnego logarytmu oszacowanej funkcji przeżycia, równoległość linii dla różnych poziomów zmiennych jakościowych świadczy o spełnieniu założenia proporcjonalnych hazardów dla danej zmiennej jakościowej. Oznacza to, że wpływ danej zmiennej jakościowej na hazard jest stały w czasie dla wszystkich obserwowanych jednostek.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Tabela 18: Weryfikacja założenia proporcjonalnych hazardów za pomocą metody graficznej*

Na podstawie wykresów ujemnego logarytmu z ujemnego logarytmu oszacowanej funkcji przeżycia, zaobserwowano brak równoległości krzywych dla niektórych zmiennych jakościowych. Ta obserwacja sugeruje potencjalne naruszenie założenia proporcjonalnych hazardów (PH) w analizowanym modelu regresji Coxa. Z tego powodu zastosowano jeszcze inną metodę weryfikacji założenia proporcjonalnych hazardów, a mianowicie reszty Shoenfelda.

## **Weryfikacja założenia proporcjonalnych hazardów – reszty Schoenfelda**

Aby zweryfikować założenie proporcjonalnych hazardów, tworzony jest wykres reszt Schoenfelda względem czasu oraz pokazany będzie współczynnik korelacji Pearsona.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

*Tabela 19: Weryfikacja założenia proporcjonalnych hazardów za pomocą metody graficzne*

Łącząc wyniki analizy graficznej z wynikami badania wykresów reszt Schoenfelda względem

czasu, można wysnuć istotne wnioski dotyczące związku pomiędzy czasem a resztami dla części zmiennych objaśniających. Obserwowana zależność niektórych zmiennych jest dodatkowo potwierdzona przez wysokie wartości korelacji Pearsona i ich istotność statystyczną. Wpływ zmiennych*, Three\_Gene\_classifier\_subtype, Lymph\_nodes\_examined\_positive\_ i Hormone\_Therapy* na czas przeżycia w przypadku raka piersi jest zależny od czasu.

Założenie proporcjonalności hazardów nie zostało spełnione w przypadku wszystkich zmiennych, postanowiliśmy w celach dydaktycznych zbudować model nieproporcjonalnych hazardów poniżej.

## **Weryfikacja założenia proporcjonalnych hazardów – zmienne zależne od czasu**

Z przeprowadzonej analizy wynika, że interakcja między pewnymi zmiennymi objaśniającymi a czasem jest statystycznie istotna. Oznacza to, że wpływ tych zmiennych na ryzyko wystąpienia badanego zdarzenia zmienia się w czasie. W rezultacie założenie o stałym wpływie tych zmiennych na ryzyko w czasie nie jest spełnione.

Dla zmiennych *Hormone\_Therapy*, *Lymph\_nodes\_examined*, *Three\_Gene\_classifier,* oraz Lymph\_nodes\_positive założenie proporcjonalnych hazardów zostało naruszone. Wyniki sugerują, że wpływ tych zmiennych na ryzyko zdarzenia nie jest stały w czasie i może zmieniać się w trakcie obserwacji.

Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

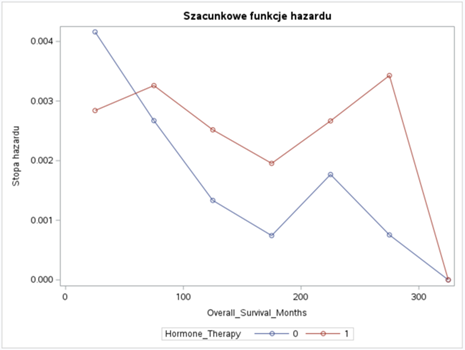
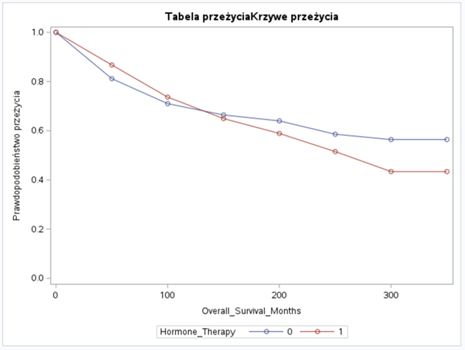
Opis wygenerowany automatycznie

*Tabela 20: Tabela istotności zmiennych z interakcjami*

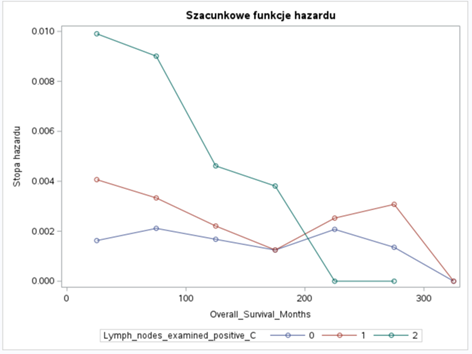
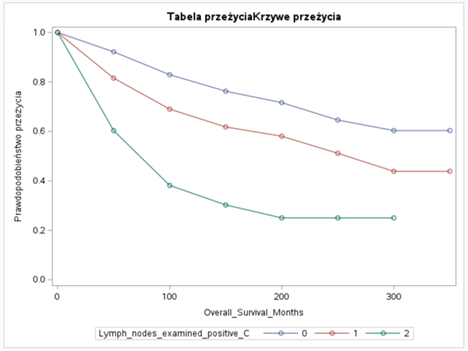
W przypadku zmiennej Hormone\_Therapy, interakcja z czasem jest statystycznie istotna, co oznacza, że wpływ hormonoterapii na ryzyko wystąpienia zdarzenia zmienia się w czasie. W związku z tym, założenie proporcjonalnych hazardów dla tej zmiennej nie jest spełnione. Podobna sytuacja ma miejsce dla zmiennej Lymph\_nodes\_examined, która również wykazuje istotną interakcję z czasem. Oznacza to, że liczba badanych węzłów chłonnych wpływa na ryzyko zdarzenia w sposób zależny od czasu, co również narusza założenie proporcjonalnych hazardów. Dla zmiennej Three\_Gene\_classifier wyniki analizy wskazują na statystycznie istotną interakcję z czasem, sugerując, że wpływ klasyfikatora trzech genów na ryzyko zdarzenia zmienia się w czasie, co również narusza założenie proporcjonalnych hazardów.

## **Modele proporcjonalnych i nieproporcjonalnych hazardów**

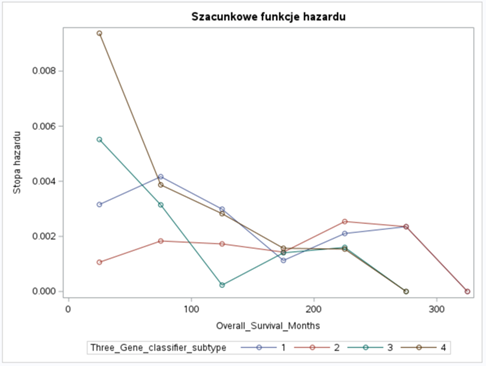
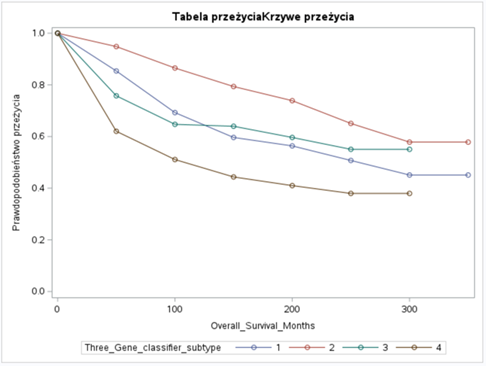
W kontekście analizy danych stwierdzono, że dla kilku zmiennych objaśniających nie zostało spełnione założenie proporcjonalnych hazardów, co sugeruje, że wpływ egzogenicznych zmiennych na ryzyko wystąpienia badanego zdarzenia może się zmieniać w zależności od czasu. W odpowiedzi na te ustalenia opracowany został model nieproporcjonalnych hazardów - wykresy funkcji przeżycia oraz funkcji hazardu.



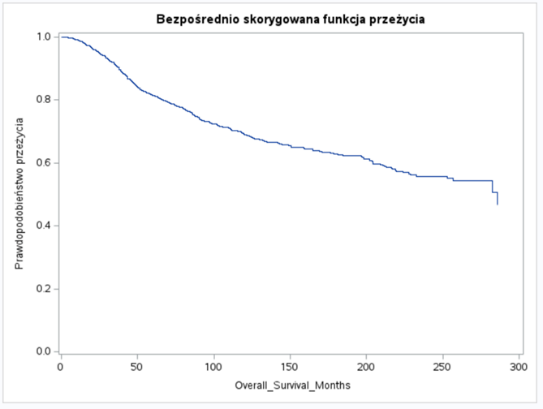
*Rysunek 11: Funkcje dla Hormone\_Therapy*



*Rysunek 12: Funkcje dla Lympt\_nodes\_examined\_positive\_C*



*Rysunek 13: Funkcje dla Three\_Gene\_Classifier*

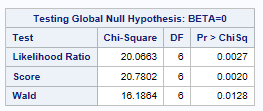


*Rysunek 14: Bezpośrednia skorygowana funkcja przeżycia*

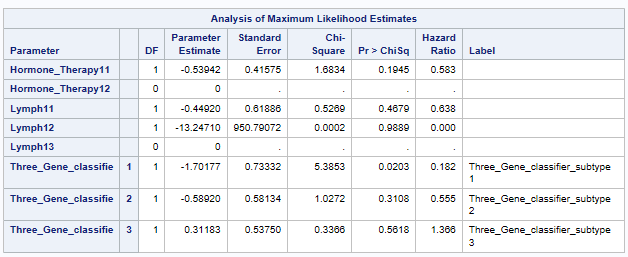
Jak więc można zauważyć, prawdopodobieństwo, że śmierć pacjentki nastąpi później niż w momencie t maleje wraz z czasem. W okresie początkowym przyjmuje ono wartość 1, bowiem wszystkie osoby uwzględnione w próbie żyją, a w miarę wymieralności pacjentek w czasie wartość funkcji przeżycia stopniowo zbliża się do osiągnięcia poziomu 0.

Dla zmiennych przyjęto następujące punkty podziałowe:

* Hormone\_Therapy: punkt podziałowy 70 (Hormone\_Therapy11, Hormone\_Therapy12)
* Lymph\_nodes\_examined: punkt podział 6, 17 (Lymth11, Lymth12, Lymth13)

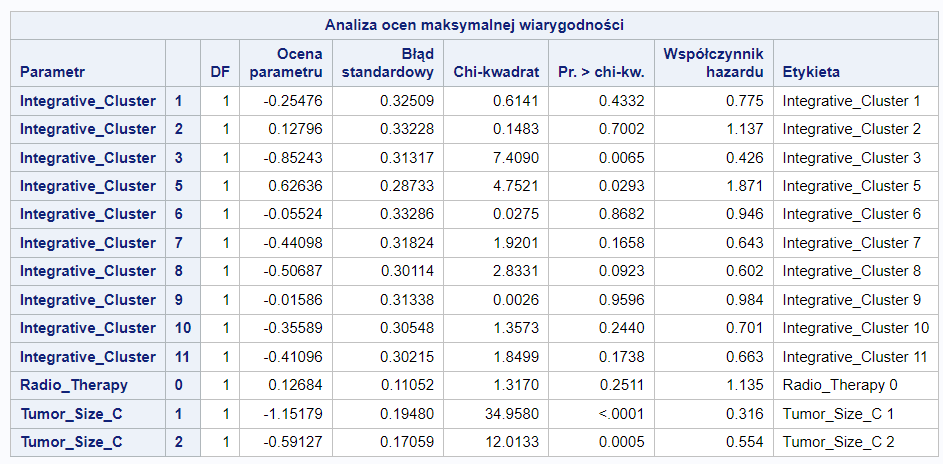
  
*Tabela 21: Testowanie hipotezy globalnej*

W pierwszym etapie weryfikowana jest hipoteza zerowa globalna. W rozważanym przykładzie wszystkie trzy testy wskazują na odrzucenie hipotezy zerowej. Oznacza to, że przynajmniej jedna zmienna ma statystycznie istotny wpływ na zmienną zależną. Poniżej jest przedstawiony model nieproporcjonalnych hazardów.



*Tabela 22: Tabela oszacowania parametrów modelu semiparametrycznego nieproporcjonalnych hazardów*

Niestety, w tym przypadku wybrane przez nas punkty przedziałowe nie okazały się istotne. Poniżej również został przedstawiony model proporcjonalnych hazardów:



Zinterpretowaliśmy wybrane istotne zmienne objaśniające w modelu proporcjonalnych hazardów. Ryzyko śmierci osoby chorej na raka piersi u której zmienna Integrative\_cluster należy do trzeciej grupy jest 0.426 razy większe niż dla grupy czwartej. Ryzyko śmierci osoby chorej na raka piersi u której zmienna Integrative\_cluster należy do piątej grupy jest 1.871 razy większe niż dla kategorii 4. Ryzyko śmierci osoby chorej na raka piersi u której zmienna Integrative\_cluster należy do trzeciej grupy jest 0.426 razy większe niż dla kategorii 4. Ryzyko śmierci osoby chorej na raka piersi u której zmienna Tumor\_Size\_C należy do pierwszej grupy jest 0.316 razy większe niż dla kategorii 3. Ryzyko śmierci osoby chorej na raka piersi u której zmienna Tumor\_Size\_C należy do pierwszej grupy jest 0.554 razy większe niż dla kategorii 3.

# **Modele semiparametryczne, podejście bayesowskie**

Podejście bayesowskie do modelowania czasu trwania opiera się na założeniu, że parametry modelu są zmiennymi losowymi o zadanym rozkładzie prawdopodobieństwa. Może być ono przydatne w przypadku, gdy istotne jest włączenie do modelu informacji pochodzących z wiedzy historycznej bądź doświadczenia i przekonań badacza. Zdecydowano się najpierw na skonstruowanie modelu o nieinformacyjnych rozkładach a priori. Podobnie jak w przypadku podejścia klasycznego, do wyboru zmiennych skorzystano z selekcji stepwise, która w porównaniu z innymi metodami, umożliwia dodatkowo wielokrotne usuwanie i włączanie zmiennych. Ostatecznymi zmiennymi w modelu okazały się *Integrative\_Cluster, Radio\_Therapy, Tumor\_Size\_C, Hormone\_Therapy, Lymph\_nodes\_examined\_positive\_C* oraz *Three\_Gene\_classifier\_subtype.* Z poprzednich analiz wiemy, że wpływ zmiennych *Hormone\_Therapy, Lymph\_nodes\_examined\_positive\_C*  i *Three\_Gene\_classifier\_subtype* na czas przeżycia w przypadku raka piersi jest zależny od czasu. Z tego powodu w modelu zawarto również trzy interakcje: *Hormone\_Therapy\_W, Lymph\_nodes\_positive\_W* i *Three\_Gene\_W*, zdefiniowane jako interakcje tych zmiennych odpowiednio z czasem mniejszym niż 80, 150 i 60 miesięcy. Istotne statystyczne punkty podziałowe zostały wyznaczone w poprzednich etapach badania. Wykorzystano 100 000 iteracji, z czego 4 000 pierwszych stanowiło cykl wstępny, aby zredukować wpływ wyboru początkowych parametrów na oszacowania. Dla wszystkich parametrów przyjęto nieinformacyjne rozkłady normalne ze średnią równą 0 i bardzo dużą wariancją (106) tak, aby przyjęcie dowolnej wartości parametru było podobnie prawdopodobne. Wartości początkowe do próbkowania dla pierwszego łańcucha Markowa oszacowano metodą częściowej wiarygodności. W pierwszej kolejności przyjrzano się przedziałom a posteriori parametrów. W badanym modelu dla 13 parametrów przedział HPD nie obejmował wartości 0, co prowadzi do następujących wniosków:

* wystąpienie skupiska komórek raka typu 5 opóźniało moment zgonu w porównaniu do skupiska typu 4ER-,
* pacjenci, którzy nie podjęli się radioterapii umierali później niż ci, którzy wzięli udział w leczeniu,
* guzy o wielkości większej bądź równej 50 mm przyczyniały się do wcześniejszych zgonów niż te o wielkości mniejszej niż 20 mm.
* brak terapii hormonalnej, opóźniał moment śmierci pacjentek posiadających raka piersi,
* pacjentki, u których wykryto obecność komórek rakowych w mniej niż 5 węzłach chłonnych żyły dłużej niż te u których wykryto ją w 5 lub więcej węzłach,
* osoby, którym przypisana została trzygenowa klasyfikacja typu ER+/HER2- High Prolif, ER+/HER2- Low Prolif oraz ER-/HER2- żyły dłużej niż te z klasyfikacją HER2+.

Ponadto istotne okazały się wszystkie interakcje. To prowadzi do wniosków, że zmienna *Hormone\_Therapy* miała bardziej pozytywny wpływ na przeżycie do 80 miesiąca przeżycia, zmienna *Lymph\_nodes\_examined\_positive\_C*  miała bardziej pozytywny wpływ na przeżycie do 150 miesiąca przeżycia, natomiast zmienna *Three\_Gene\_classifier\_subtype* miała bardziej pozytywny wpływ na przeżycie do 60 miesiąca po diagnozie.

Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznie

*Tabela 23: Przedziały a posteriori dla semiparametrycznego nieinformacyjnego modelu bayesowskiego*

Badając autokorelacje a posteriori, możemy zauważyć, że dla wszystkich parametrów wartości autokorelacji stopniowo bezwzględnie maleją dla coraz większych opóźnień, co świadczy o prawidłowym mieszaniu łańcuchów.

Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznie

*Tabela 23: Autokorelacje a posteriori dla semiparametrycznego nieinformacyjnego modelu bayesowskiego*

Przeprowadzono testy statystyczne w celu zbadania własności wygenerowanych łańcuchów Markowa. Pierwszym testem był test Gelmana-Rubina, w którym otrzymane wartości bliskie 1 dla wszystkich parametrów oznaczają zbieżność wszystkich łańcuchów. Ponadto wysokie wartości skutecznych rozmiarów prób (ESS) i niskie czasy autokorelacji świadczą o dobrym mieszaniu wygenerowanych łańcuchów Markowa.

Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznie

*Tabela 24: Test Gelmana Rubina dla semiparametrycznego nieinformacyjnego modelu bayesowskiego*

Obraz zawierający tekst, zrzut ekranu, numer, menu

Opis wygenerowany automatycznie

*Tabela 25: Efektywne liczebności prób dla semiparametrycznego nieinformacyjnego modelu bayesowskiego*

Następnie przeprowadzono test Gewek’a w którym na podstawie otrzymanych wyników dla wszystkich przypadków nie ma podstaw do odrzucenia hipotezy zerowej o zbieżności łańcuchów, dla dowolnego poziomu istotności.

Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznie

*Tabela 25: Test Geweke’a dla semiparametrycznego nieinformacyjnego modelu bayesowskiego*

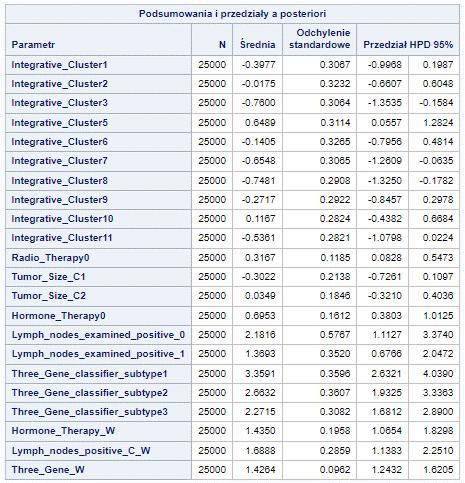
Według testu Raftery'ego-Lewisa, który wyznacza liczbę iteracji odrzuconych potrzebnych do osiągnięcia pożądanej precyzji estymacji, ilość ta nie przekracza założonej liczby 4 000 iteracji wstępnych. Ostatnim testem był test Heidelbergera-Welcha. Na podstawie testu stacjonarności wszystkie łańcuchy są zbieżne do rozkładów stacjonarnych. Test półrozstępu zweryfikował, że dla większości parametrów liczność próby a posteriori jest wystarczająca do estymacji wartości średniej przy poziomie istotności 0,05, z wyłączeniem niektórych kategorii zmiennej *Integratice\_Cluster*. Na podstawie testów możemy wnioskować o poprawnych własnościach łańcuchach Markowa w modelu.

Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznie

*Tabela 26: Test Heidelgergera-Welcha dla semiparametrycznego nieinformacyjnego modelu bayesowskiego*

Przygotowano także informacyjny model bayesowski, w którym zakłada się pewne informacyjne rozkłady a priori parametrów. W pierwszym kroku wylosowano 200 obserwacji z całego zbioru. Wszystkie modele bayesowskie opierały się na pozostałej części obserwacji. Wylosowana próba została użyta jako historyczna informacja będąca podstawą rozkładów a priori parametrów. Zbudowano na niej model, na takich samych zmiennych jak w przypadku modelu nie informacyjnego, w którym wykorzystano 100 000 iteracji i 4 000 cykli wstępnych („spalonych”). Oszacowania parametrów w tym modelu zostały wykorzystane jako średnie i odchylenia standardowe rozkładów a priori parametrów w nowym modelu informacyjnym. Zbudowany model również miał łącznie 100 000 iteracji i 4 000 iteracji początkowych.



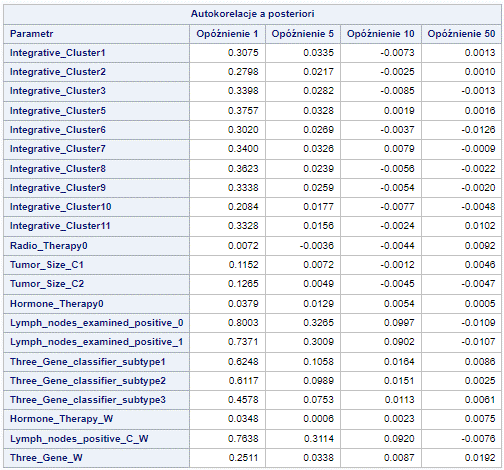
*Tabela 27: Przedziały a posteriori dla semiparametrycznego informacyjnego modelu bayesowskiego*

Na podstawie przedziałów a posteriori możemy zaobserwować, że zmienna *Tumor\_Size\_C* stała się nieistotna, natomiast 3 poziomy zmiennej Integrative\_Cluster uzyskały istotność statystyczną w modelu. Prowadzi to do dodatkowych wniosków:

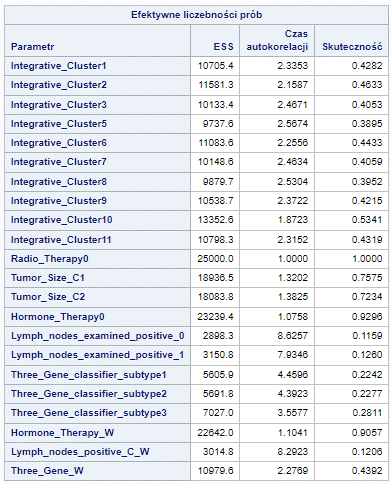
- wystąpienie skupiska komórek raka typu 3 przyśpieszało moment zgonu w porównaniu do skupiska typu 4ER-,

- wystąpienie skupiska komórek raka typu 7 i 8 przyśpieszało moment zgonu w porównaniu do skupiska typu 4ER-.

Na podstawie autokorelacji aposteriori uznajemy, że w modelu informacyjnym również możemy wnioskować o dobrym mieszaniu łańcuchów. Wnioski te potwierdzone są dużymi wartościami efektywnych liczebności prób dla wszystkich parametrów.

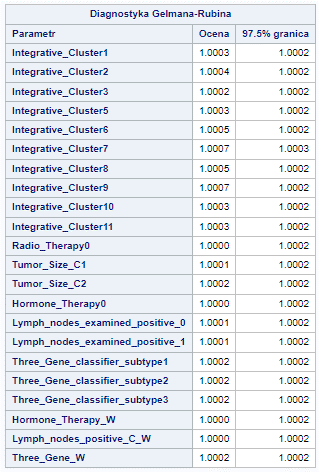


*Tabela 28: Autokorelacje a posteriori dla semiparametrycznego informacyjnego modelu bayesowskiego*



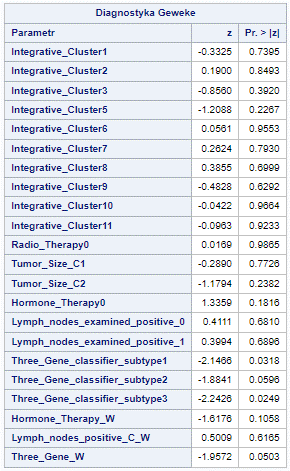
*Tabela 29: Efektywne liczebności prób dla semiparametrycznego informacyjnego modelu bayesowskiego*

Wyniki testu Gelmana-Rubina nie zmieniły się – w tym modelu również uznajemy zbieżność wszystkich łańcuchów. Podobne wnioski mamy także w przypadku badania efektywnych liczebności prób - posiadają duże wartości więc występuje dobre mieszanie łańcuchów Markowa.



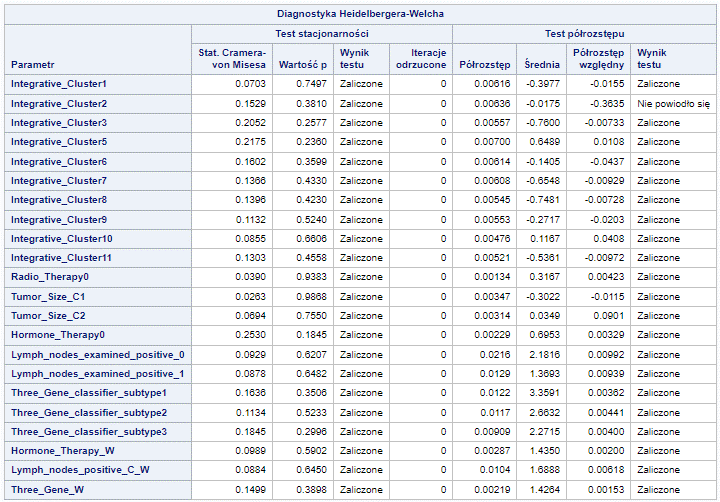
*Tabela 30: Test Gelmana Rubina dla semiparametrycznego informacyjnego modelu bayesowskiego*

Następnie wykonano test Gewek’a w którym dla więszości parametrów nie ma podstaw do odrzucenia hipotezy zerowej o zbieżności łańcuchów, dla dowolnego poziomu istotności, oprócz zmiennej *Lymph\_nodes\_examined\_positive*.



*Tabela 31: Test Geweke’a dla semiparametrycznego informacyjnego modelu bayesowskiego*

Według testu Raftery'ego-Lewisa liczba iteracji odrzuconych potrzebnych do osiągnięcia pożądanej precyzji estymacji nadal nie przekracza założonej liczby 4 000 iteracji wstępnych. Analizę modelu informacyjnego zakończono testem Heidelbergera-Welcha. Na podstawie testu stacjonarności wszystkie łańcuchy nadal są zbieżne do rozkładów stacjonarnych. W teście półrozstępu jedynie w przypadku kategorii *Integrative\_Cluster2* liczność próby a posteriori nie jest wystarczająca do estymacji wartości średniej przy poziomie istotności 0,05. W przypadku tego modelu informacyjnego możemy wnioskować o dobrych własnościach łańcuchach Markowa w modelu.



*Tabela 31: Test Heidelgergera-Welcha dla semiparametrycznego informacyjnego modelu bayesowskiego*

Obliczając statystyki porównawcze, takie jak DIC, dla modelu nieinformacyjnego i informacyjnego możemy stwierdzić, że model informacyjny jest nieco lepiej dopasowany do danych.

|  |  |  |
| --- | --- | --- |
| Kryterium | Bayesowski model nieinformacyjny | Bayesowki model informacyjny |
| DIC | 3262.367 | 3261.893 |
| pD | 22.071 | 20.104 |
| EAIC | 3306.509 | 3302.101 |
| EBIC | 18939.426 | 18936.274 |

*Tabela 32: Porównanie Bayesowskiego modelu informacyjnego i nie informacyjnego*

# **Podsumowanie**

W analizie wykorzystano modele parametryczne, nieparametryczne i semiparametryczne, przy czym zastosowano podejście klasyczne oraz bayesowskie. Najważniejszymi zmiennymi w obszarze przewidywania przeżycia okazały się *zmienne Integrative\_Cluster, Radio\_Therapy, Tumor\_Size, Hormone\_Therapy, Lymph\_nodes\_examined\_positive* oraz *Three\_Gene\_classifier\_subtype.*

W modelu wykładniczym wszystkie zmienne objaśniające okazały być istotne statystycznie, ale odrzuciliśmy hipotezę zerową o stałości funkcji hazardu i przyjęliśmy H1, co oznacza że model wykładniczy nie jest odpowiedni do opisu badanego procesu.

Analiza założenia proporcjonalności hazardów wykazała, iż zmienne *Hormone\_Therapy, Lymph\_nodes\_examined\_positive* oraz *Three\_Gene\_classifier\_subtype* posiadały zmienny w czasie wpływ na przeżycie pacjentów. Ponadto modelu semiparametrycznym klasycznym z proporcjonalnymi hazardami istotne okazały się wpływy zmiennych *Integrative\_Cluster3*, *Integrative\_Cluster5*, *Tumor\_Size1* oraz *Tumor\_Size2*.

Ostatecznie model informacyjny w ramach podejścia bayesowskiego okazał się bardziej efektywny niż model nieinformacyjny. Wyniki analizy wskazują, że wystąpienie skupiska komórek raka typu 5 opóźniało moment zgonu w porównaniu do skupiska typu 4ER-. Pacjenci, którzy nie podjęli się radioterapii, umierali później niż ci, którzy wzięli udział w leczeniu. Guzy o większej wielkości przyczyniały się do wcześniejszych zgonów. Brak terapii hormonalnej opóźniał moment śmierci pacjentek z rakiem piersi, a obecność komórek rakowych w mniej niż 5 węzłach chłonnych znacząco przedłużała przeżycie. Warto również zauważyć, że wszystkie interakcje okazały się istotne w zbudowanych modelach.

# **Spis tabel i rysunków**

**Tabele**

Tabela 1: Szczegóły zmiennej cenzurującej

Tabela 2: Szczegóły zmiennej objaśnianej

Tabela 3: Szczegóły zmiennych objaśniających

Tabela 4: Kategoryzacja zmiennych objaśniających

Tabela 5: Wartości tabeli przeżycia opracowane metodą tradycyjną

Tabela 6: Podsumowanie wartości ocenzurowanych i nieocenzurowanych

Tabela 7: Testy dla zmiennych Age\_at\_Diagnosis\_c, Hormone\_Therapy, Radio\_Therapy

Tabela 8: Testy dla zmiennych Chemotherapy, Cancer\_Type\_Detailed, Cellularity

Tabela 9: Testy dla zmiennych Tumor\_size\_c i Three\_Gene\_classifier\_subtype

Tabela 10: Wartości tabeli przeżycia opracowane metodą Kaplana-Meiera

Tabela 11: Analiza efektów typu III dla modelu wykładniczego

Tabela 12: Test mnożnika Lagrange'a dla modelu wykładniczego

Tabela 13: Informacje o modelu wykładniczym

Tabela 14: Statystyki dopasowania modelu wykładniczego dla zmiennej zależnej Overall\_Survival\_Months

Tabela 15: Statystyki dopasowania modelu wykładniczego dla odpowiedzi nierejestrowanej

Tabela 16: Analiza ocen parametrów maksymalnej wiarygodności

Tabela 17: Analiza ocen parametrów maksymalnej wiarygodności

Tabela 18: Weryfikacja założenia proporcjonalnych hazardów za pomocą metody graficznej

Tabela 19: Weryfikacja założenia proporcjonalnych hazardów za pomocą metody graficzne

Tabela 20: Tabela ustalania istotności interakcji

Tabela 21: Testowanie hipotezy globalnej

Tabela 22: Tabela oszacowania parametrów modelu semiparametrycznego nieproporcjonalnych hazardów

Tabela 23: Autokorelacje a posteriori dla semiparametrycznego nieinformacyjnego modelu bayesowskiego

Tabela 24: Test Gelmana Rubina dla semiparametrycznego nieinformacyjnego modelu bayesowskiego

Tabela 25: Test Geweke’a dla semiparametrycznego nieinformacyjnego modelu bayesowskiego

Tabela 26: Test Heidelgergera-Welcha dla semiparametrycznego nieinformacyjnego modelu bayesowskiego

Tabela 27: Przedziały a posteriori dla semiparametrycznego informacyjnego modelu bayesowskiego

Tabela 28: Autokorelacje a posteriori dla semiparametrycznego informacyjnego modelu bayesowskiego

Tabela 29: Efektywne liczebności prób dla semiparametrycznego informacyjnego modelu bayesowskiego

Tabela 30: Test Gelmana Rubina dla semiparametrycznego informacyjnego modelu bayesowskiego

Tabela 31: Test Geweke’a dla semiparametrycznego informacyjnego modelu bayesowskiego

Tabela 32: Porównanie Bayesowskiego modelu informacyjnego i nie informacyjnego

**Rysunki**

Rysunek 1: Histogram zmiennej objaśnianej

Rysunek 2: Histogramy zmiennych objaśniających

Rysunek 3: Krzywa przeżycia dla zmiennej Overall\_Survival\_Months

Rysunek 4: Funkcja gęstości prawdopodobieństwa dla zmiennej Overall\_Survival\_Months

Rysunek 5: Funkcja hazardu dla zmiennej Overall\_Survival\_Months

Rysunek 6: Krzywa przeżycia dla zmiennej Tumor\_size\_C

Rysunek 7: Funkcja hazardu dla zmiennej Tumor\_size\_C

Rysunek 8: Funkcja przeżycia dla zmiennej Three\_Gene\_classifier\_subtype

Rysunek 9: Funkcja hazardu dla zmiennej Three\_Gene\_classifier\_subtype

Rysunek 10: Funkcje dla metody Kaplana-Meiera

Rysunek 11: Funkcje dla Hormone\_Therapy

Rysunek 12: Funkcje dla Lympt\_nodes\_examined\_positive\_C

Rysunek 13: Funkcje dla Three\_Gene\_Classifier

Rysunek 14: Bezpośrednia skorygowana funkcja przeżycia