ANALIZA PRZEŻYCIA

SPRAWOZDANIE

OPRACOWAŁA:
ALEKSANDRA GRZESZCZUK
NUMER ALBUMU: 255707

SPIS TREŚCI

1 ZADANIE 1

Zbadamy prawdopodobieństwo przeżycia według płci. Wykorzystamy w tym celu funkcję `Surv` oraz `survfit`, które służą do oszacowania przeżycia `Kaplana - Meiera`.

```
fit_sex <- survfit(Surv(time, status) ~ sex, data = lung,
                   type = "kaplan-meier")
```

Wyniki z funkcji przedstawia poniższa tabela:

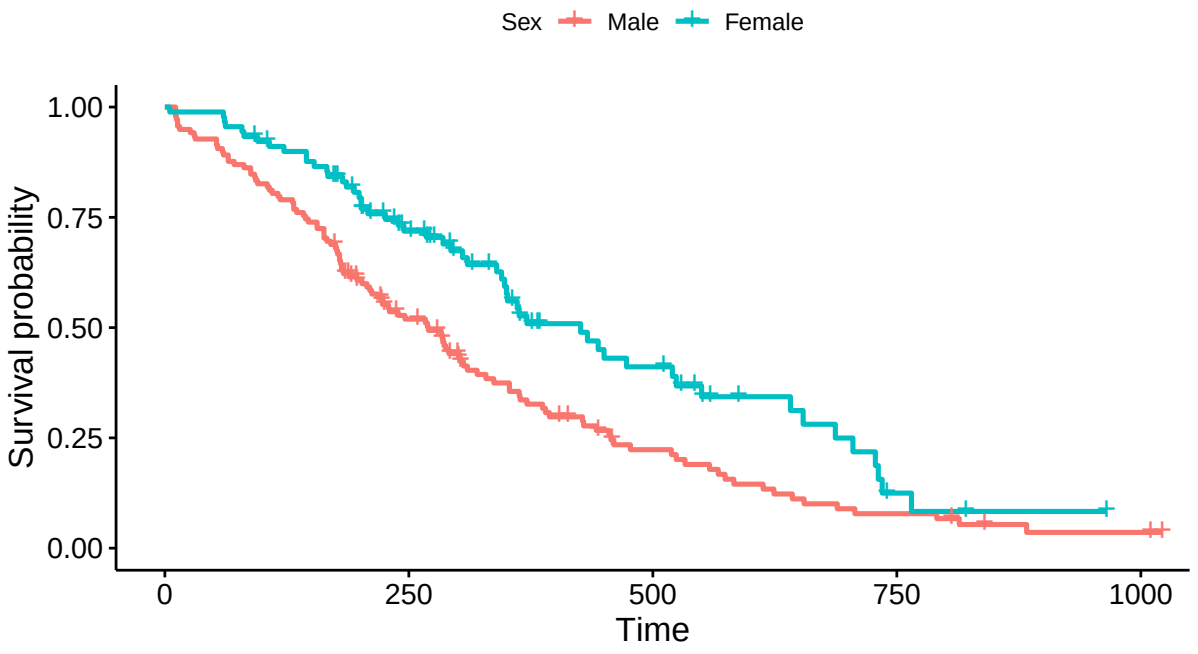
	Płeć	Liczba obserwacji	Ilość zdarzeń	Mediana przeżycia	LCL	UCL
1	Mężczyźni	138	112	270	212	310
2	Kobiety	90	53	426	348	550

Tabela 1: Krótkie podsumowanie krzywych przeżycia - opracowanie własne

Mediana przeżycia kobiet jest prawie dwa razy większa niż mediana przeżycia mężczyzn. Sugeruje to, że większa śmiertelność z powodu raka płuc panuje u mężczyzn.

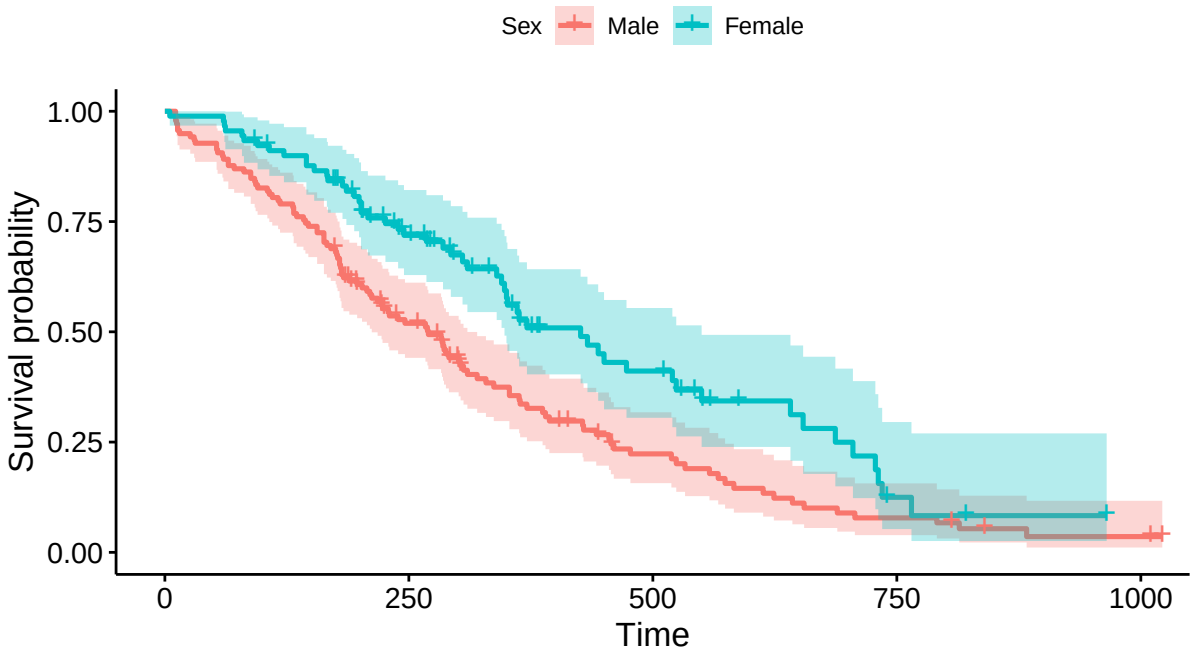
Poniższy wykres (narysowany korzystając z funkcji `ggsurvplot` z biblioteki `survminer`) potwierdza wcześniejsze wnioski. Niebieska krzywa będą estymatorem `KM` kobiet przez cały okres badania jest ponad czerwoną krzywą estymatora `KM` mężczyzn. To znaczy, że w całym okresie badania prawdopodobieństwo przeżycia kobiet jest większe. Każde zdarzenie (czyli śmierć pacjenta) jest sygnalizowane pionowym spadkiem krzywej.

```
ggsurvplot(fit_sex, legend.title = "Sex",
           legend.labs = c("Male", "Female"),)
```



Rysunek 1: Krzywa Kaplana - Meiera dla zmiennej płci bez przedziałów ufności

```
ggsurvplot(fit_sex, conf.int = TRUE, legend.title = "Sex",
           legend.labs = c("Male", "Female"))
```



Rysunek 2: Krzywa Kaplana - Meiera dla zmiennej płci z przedziałami ufności

Test `log - rank` jest dość popularnym testem stosowanym w analizie przeżycia, ponieważ może zostać użyty także w przypadku wystąpienia wartości cenzurowanych. 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.

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.

```
survdiff(Surv(time, status) ~ sex, data = lung, rho = 0)

## Call:
## survdiff(formula = Surv(time, status) ~ sex, data = lung, rho = 0)
##
##         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
```

```r
survdiff(Surv(time, status) ~ sex, data = lung, rho = 1)

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

Widzimy, że dla obydwu testów $p - value < 0.05 = \alpha$, zatem odrzucamy hipotezę zerową mówiącą o równości rozkładów przeżycia ze względu na płeć i przyjmujemy hipotezę alternatywną - przeżycie jest funkcją zależną od płci chorego. Dodatkowo skorzystamy jeszcze z nieparametrycznego testu `Coxa` (funkcja `coxph`) (zwanego inaczej modelem proporcjonalnych hazardów).

```r
coxph(Surv(time, status) ~ sex, data = lung)

## 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
```

Ponownie otrzymaliśmy $p - value = 0.001111 < 0.05 = \alpha$, zatem przyjmujemy hipotezę alternatywną o nierówności rozkładów przeżycia.

2 ZADANIE 2

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. Dokonamy teraz kategoryzacji tej zmiennej na 3 podgrupy:

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

```r
wiek <- cut(lung$age, breaks = c(0, 50, 65, Inf))
```

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

Tabela 2: Kategoryzacja zmiennej wiek

Korzystając z funkcji `Surv` oraz `survfit` dokonamy krótkiego podsumowania kategoryzowanej zmiennej wiek a następnie wyznaczymy wykres estymatora `Kaplana - Meiera` dzięki funkcji `ggsurvplot`.

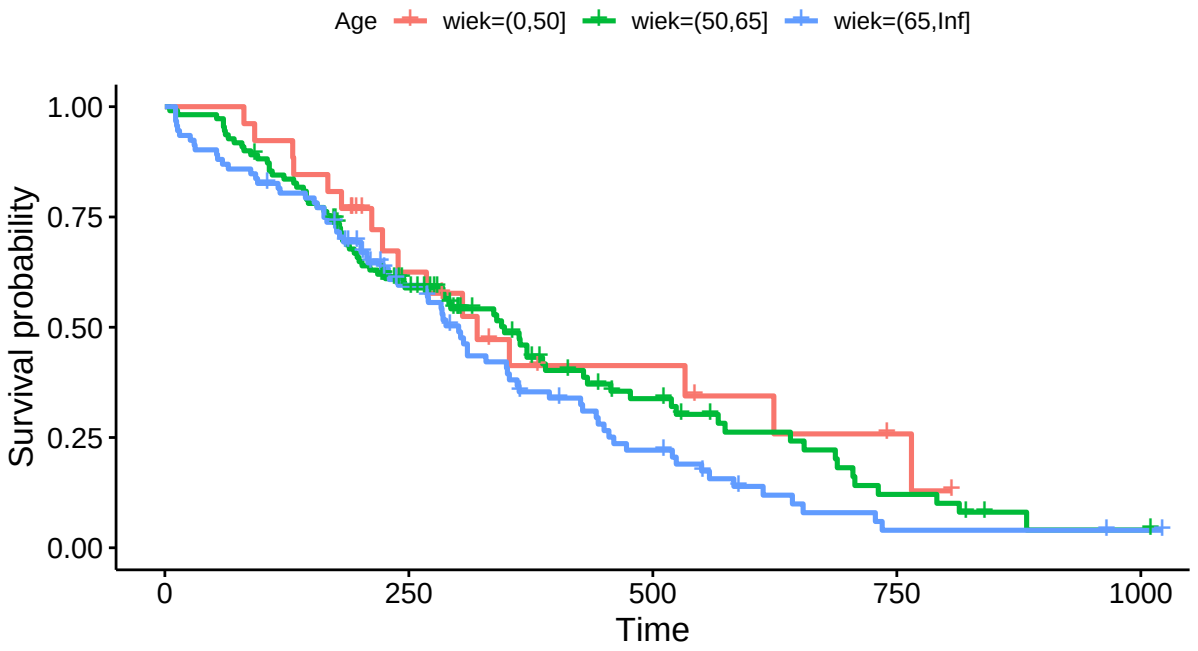
```
fit_age <- survfit(Surv(time, status) ~ wiek, data = lung,
                   type = "kaplan-meier")
```

Wyniki z funkcji przedstawia poniższa tabela:

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

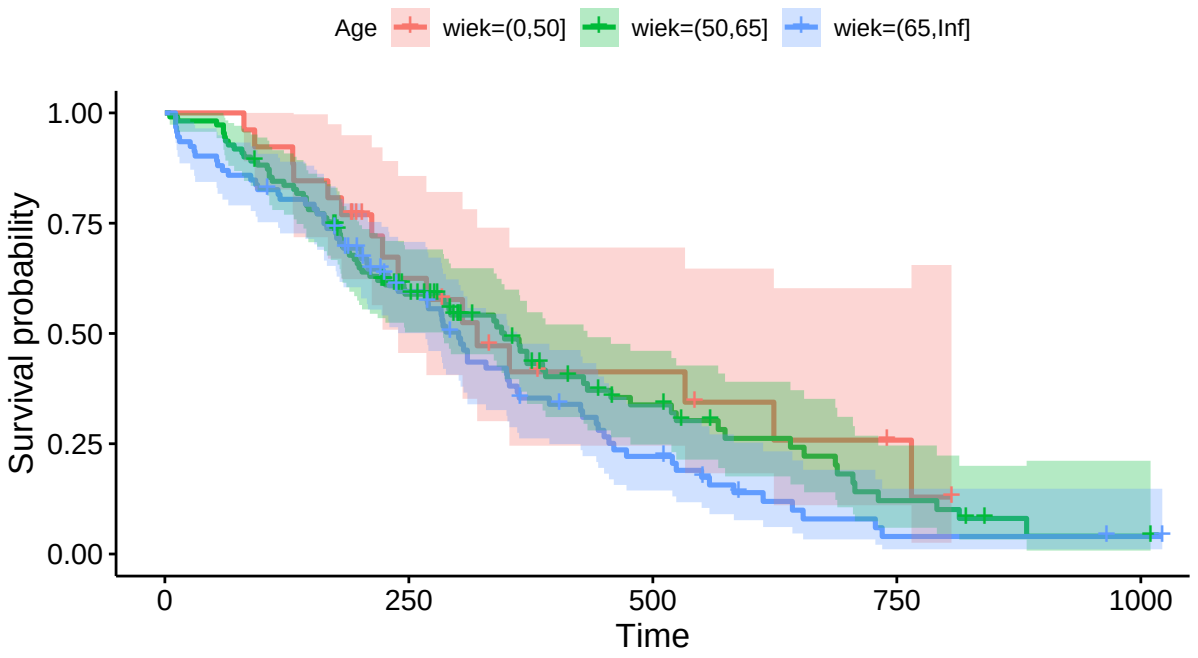
Tabela 3: Krótkie podsumowanie kategorii wiekowych - opracowanie własne

```
ggsurvplot(fit_age, legend.title = "Age")
```



Rysunek 3: Krzywa Kaplana - Meiera dla zmiennej kategoryzowanej wiek bez przedziałów ufności

```r
ggsurvplot(fit_age, conf.int = TRUE, legend.title = "Age")
```



Rysunek 4: Krzywa Kaplana - Meiera dla zmiennej kategoryzowanej wiek z przedziałami ufności

Powyżej przedstawiony został estymator funkcji przeżycia `Kaplana - Meiera` dla pacjentów z rakiem płuc z podziałem na wiek. Widzimy znacznie spadającą funkcję schodkową, co potwierdza nasze wcześniejsze wnioski, że bez względu na wiek pacjenta, szanse na przeżycie są małe.

Tak jak w poprzednim zadaniu, skorzystamy z funkcji `log - rank` dla róznych poziomów wartości `rho`.

```r
survdiff(Surv(time, status) ~ wiek, data = lung, rho = 0)
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ wiek, data = lung, rho = 0)
##
##                 N Observed Expected (O-E)^2/E (O-E)^2/V
## wiek=(0,50]    26       16     20.6     1.019     1.175
## wiek=(50,65]  110       76     81.9     0.424     0.847
## wiek=(65,Inf]  92       73     62.5     1.753     2.855
##
##  Chisq= 3.2  on 2 degrees of freedom, p= 0.2
```

```r
survdiff(Surv(time, status) ~ wiek, data = lung, rho = 1)
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ wiek, data = lung, rho = 1)
##
##                 N Observed Expected (O-E)^2/E (O-E)^2/V
## wiek=(0,50]    26     9.45     12.1     0.563     0.907
## wiek=(50,65]  110    45.41     48.4     0.187     0.514
## wiek=(65,Inf]  92    44.25     38.6     0.815     1.877
##
##  Chisq= 2.2  on 2 degrees of freedom, p= 0.3
```

Widzimy, że dla obydwu testów $p-value > 0.05 = \alpha$ zatem przyjmujemy hipotezę zerową mówiącą o równości rozkładów przeżycia dla trzech wybranych przeze mnie kategorii wiekowych. Dodatkowo skorzystamy jeszcze z nieparametrycznego testu `Coxa` (funkcja `coxph`)(zwanego inaczej modelem proporcjonalnych hazardów).

```
coxph(Surv(time, status) ~ wiek, data = lung)

## Call:
## coxph(formula = Surv(time, status) ~ wiek, data = lung)
##
##                coef exp(coef) se(coef)     z     p
## wiek(50,65]  0.1791    1.1961   0.2756 0.650 0.516
## wiek(65,Inf] 0.4114    1.5089   0.2771 1.484 0.138
##
## Likelihood ratio test=3.25  on 2 df, p=0.1974
## n= 228, number of events= 165
```

Ponownie otrzymaliśmy $p - value = 0.1974 > 005 = \alpha$, zatem przyjmuejmy hipotezę o równości przeżycia rozkładów.

Analizując powyższe wnioski, otrzymaliśmy, że w momencie zachorowania na raka płuc, właściwie bez względu na wiek, nasze szanse na przeżycie są dość małe.

Dokonamy jeszcze innej kategoryzacji zmiennej wiek.

```
wiek1 <- cut(lung$age, breaks = c(0, 20, 40, 60, 80, Inf))
```

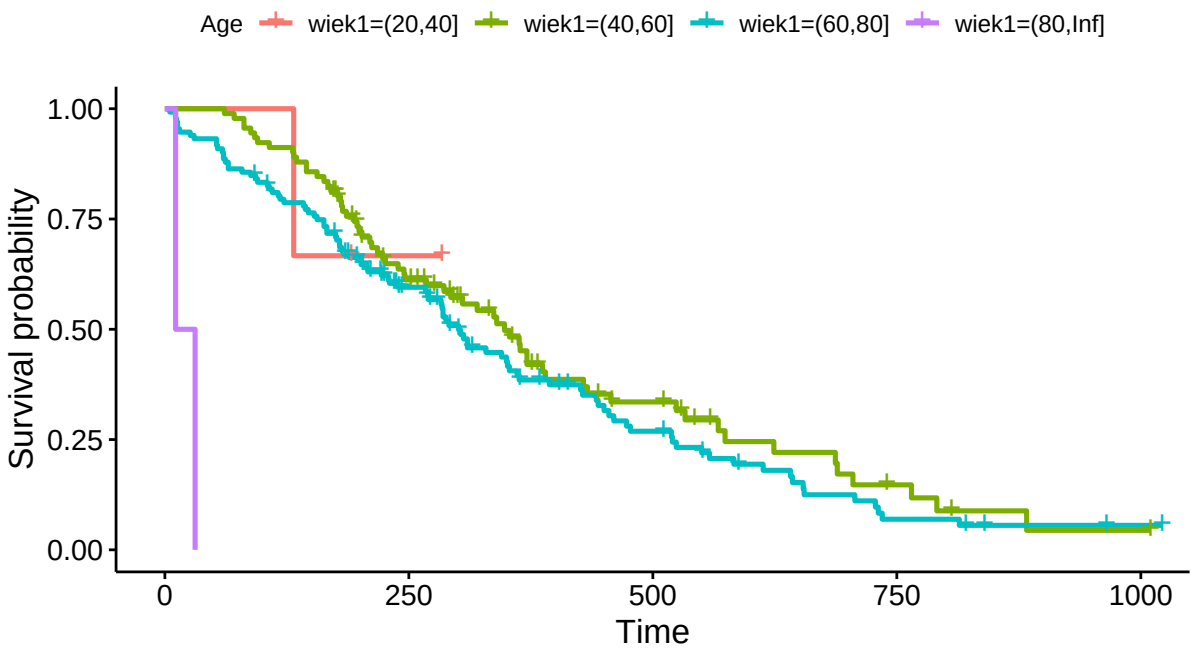
```
fit_age1 <- survfit(Surv(time, status) ~ wiek1, data = lung,
            type = "kaplan-meier")
print(fit_age1)

## Call: survfit(formula = Surv(time, status) ~ wiek1, data = lung, type = "kaplan-mei
##
##                  n events median 0.95LCL 0.95UCL
## wiek1=(20,40]    3      1     NA     132      NA
## wiek1=(40,60]   91     63    348     286     433
## wiek1=(60,80]  132     99    301     269     361
## wiek1=(80,Inf]   2      2     21      11      NA
```

Widzimy, że najwięcej zdarzeń zaszło nam w przedziale wiekowym $(60, 80]$, zaś najmniej w przedziale $(20, 40]$. Jednakże patrząc procentowo, otrzymaliśmy 100% zdarzeń w ostatnim przedziale wiekowym - osoby od 80 lat.

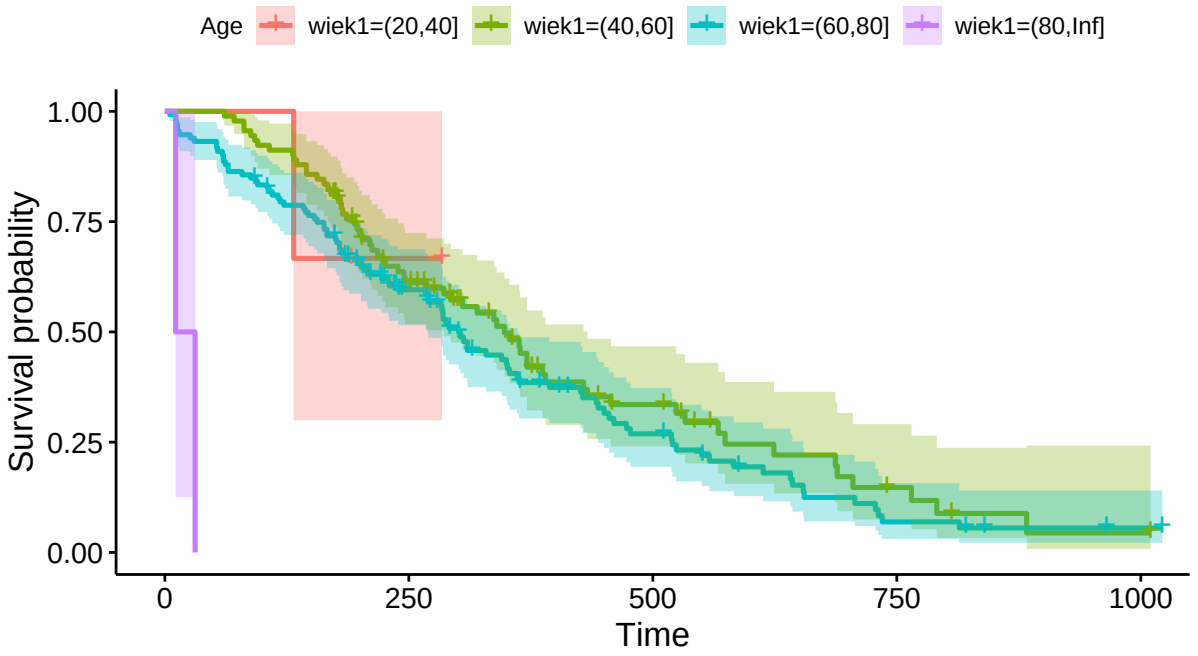
Rysujemy wykresy estymatora `Kaplana - Meiera` bez przedziałów ufności oraz z przedziałami.

```
ggsurvplot(fit_age1, legend.title = "Age")
```



Rysunek 5: Krzywa Kaplana - Meiera dla zmiennej kategoryzowanej wiek bez przedziałów ufności

```
ggsurvplot(fit_age1, conf.int = TRUE, legend.title = "Age")
```



Rysunek 6: Krzywa Kaplana - Meiera dla zmiennej kategoryzowanej wiek z przedziałami ufności

Oraz korzystając ze wcześniejszych funkcji, sprawdzamy hipotezę zerową mówiącą o równości rozkładów czasu życia w podgupach ze względu na wiek.

```
survdiff(Surv(time, status) ~ wiek1, data = lung, rho = 0)

## Call:
## survdiff(formula = Surv(time, status) ~ wiek1, data = lung, rho = 0)
##
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## wiek1=(20,40]    3        1   1.1311    0.0152    0.0154
## wiek1=(40,60]   91       63  71.4857    1.0073    1.7852
## wiek1=(60,80]  132       99  92.3163    0.4839    1.1032
## wiek1=(80,Inf]   2        2   0.0669   55.8931   56.5059
##
##  Chisq= 58  on 3 degrees of freedom, p= 2e-12
```

```
survdiff(Surv(time, status) ~ wiek1, data = lung, rho = 1)

## Call:
## survdiff(formula = Surv(time, status) ~ wiek1, data = lung, rho = 1)
##
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## wiek1=(20,40]    3    0.829   0.9217   0.00942    0.0115
## wiek1=(40,60]   91   35.682  42.7508   1.16899    2.8967
## wiek1=(60,80]  132   60.644  55.3679   0.50285    1.6035
## wiek1=(80,Inf]   2    1.952   0.0658  54.06433   55.5413
```

```
##
##  Chisq= 57.9  on 3 degrees of freedom, p= 2e-12
```

W obydwu przypadkach, $p - value = 2e - 12 < 0.05 = \alpha$, zatem odrzucamy hipotezę zerową mówiącą o równości rozkładów i przyjmujemy hipotezę alternatywną zakładającą, że przeżycie zależy od wieku.

```
coxph(Surv(time, status) ~ wiek1, data = lung)

## Call:
## coxph(formula = Surv(time, status) ~ wiek1, data = lung)
##
##                    coef exp(coef) se(coef)      z        p
## wiek1(20,40]  -3.77763   0.02288  1.26717 -2.981  0.00287
## wiek1(40,60]  -3.79416   0.02250  0.79325 -4.783 1.73e-06
## wiek1(60,80]  -3.59627   0.02743  0.78893 -4.558 5.15e-06
## wiek1(80,Inf]       NA        NA  0.00000     NA       NA
##
## Likelihood ratio test=11.91  on 3 df, p=0.00768
## n= 228, number of events= 165
```

Test `Coxa` również potwierdza nasze przypuszczenia. ($p - vaue = 0.00768 < 0.05 = \alpha$). Wnioskujemy zatem, że rodzaj kategoryzacji zmiennej wiek może mieć wpływ na uzyskiwane przez nas wyniki.

3 ZADANIE 3

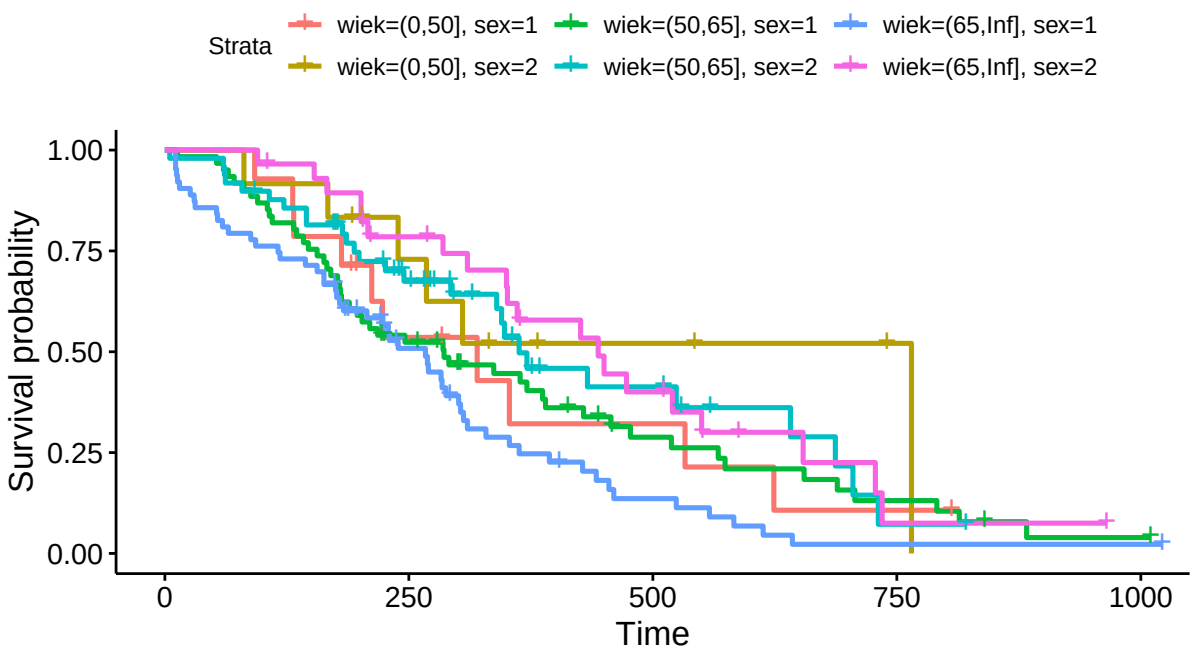
Sprawdzimy, czy uwzględnienie zarówno zmiennej płci jak i kategoryzowanej zmiennej wiek, wpływa na hipotezę o równości rozkładów przeżycia, czy też nie.
Dokonujemy krótkiego podsumowania, którego wyniki przedstawia poniższa tabela:

```
fit_sex_age <- survfit(Surv(time, status) ~ wiek + sex, data = lung,
                       type = "kaplan-meier")
```

Płeć	Przedział	Liczba obserwacji	Ilość zdarzeń	Mediana
Kobiety	$(0, 50]$	12	6	765
Kobiety	$[50, 65)$	49	27	363
Kobiety	$[65, 100]$	29	20	444
Mężczyźni	$(0, 50]$	14	10	320
Mężczyźni	$[50, 65)$	61	49	286
Mężczyźni	$[65, 100]$	63	53	267

Tabela 4: Krótkie podsumowanie kategorii wiekowych - opracowanie własne

```r
ggsurvplot(fit_sex_age)
```




```r
survdiff(Surv(time, status) ~ wiek + sex, data = lung, rho = 0)
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ wiek + sex, data = lung,
##     rho = 0)
##
##                         N Observed Expected (O-E)^2/E (O-E)^2/V
## wiek=(0,50], sex=1     14       10     10.1  2.78e-04  2.98e-04
## wiek=(0,50], sex=2     12        6     10.5  1.95e+00  2.10e+00
## wiek=(50,65], sex=1    61       49     47.2  6.69e-02  9.53e-02
## wiek=(50,65], sex=2    49       27     34.7  1.70e+00  2.16e+00
## wiek=(65,Inf], sex=1   63       53     34.3  1.02e+01  1.31e+01
## wiek=(65,Inf], sex=2   29       20     28.2  2.40e+00  2.92e+00
##
##  Chisq= 16.6  on 5 degrees of freedom, p= 0.005
```

```r
survdiff(Surv(time, status) ~ wiek + sex, data = lung, rho = 1)
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ wiek + sex, data = lung,
##     rho = 1)
##
##                         N Observed Expected (O-E)^2/E (O-E)^2/V
## wiek=(0,50], sex=1     14     5.96     6.08   0.00233   0.00346
## wiek=(0,50], sex=2     12     3.48     5.97   1.03606   1.58708
## wiek=(50,65], sex=1    61    29.35    26.86   0.23190   0.45222
```

```
## wiek=(50,65], sex=2  49    16.06    21.56   1.40354   2.50759
## wiek=(65,Inf], sex=1 63    35.06    22.63   6.82634  12.12891
## wiek=(65,Inf], sex=2 29     9.18    16.00   2.90513   5.08904
##
##  Chisq= 17.4  on 5 degrees of freedom, p= 0.004
```

Widzimy, że obydwa testy otrzymały $p-value$ mniejszą od poziomu istotności. Odrzucamy zatem hipotezę zerową mówiącą o równości rozkładów czasu przeżycia w tej grupie i przyjmujemy hipotezę alternatywną. To znaczy, że czas życia nie zależy od zmiennej płci połączonej z kategoryzowaną zmienną wiek.

```
coxph(Surv(time, status) ~ wiek + sex, data = lung)

## Call:
## coxph(formula = Surv(time, status) ~ wiek + sex, data = lung)
##
##                 coef exp(coef) se(coef)      z       p
## wiek(50,65]   0.1346    1.1441   0.2759  0.488 0.62557
## wiek(65,Inf]  0.3824    1.4658   0.2773  1.379 0.16788
## sex          -0.5304    0.5883   0.1675 -3.168 0.00154
##
## Likelihood ratio test=13.82  on 3 df, p=0.003163
## n= 228, number of events= 165
```

Test `Coxa` również potwierdza nasze wnioski.

Sprawdzimy jeszcze jakie wnioski otrzymaliśmy, nie kategoryzując zmiennej wieku.

```
funkcja1 <- survdiff(Surv(time, status) ~ age + sex, data = lung, rho = 0)
```

```
funkcja2 <- survdiff(Surv(time, status) ~ age + sex, data = lung, rho = 1)
```

```
funkcja3 <- coxph(Surv(time, status) ~ age + sex, data = lung)
```

Wyniki przedstawia poniższa tabela:

Funkcja	p - value
1 (log - rank)	3e-14
2 (Ghan)	8e-12
3 (Cox)	0.0008574

Tabela 5: Krótkie podsumowanie kategorii wiekowych - opracowanie własne

W każdym z przypadków $p - value$ jest mniejsza od poziomu istotności α, zatem przyjmujemy hipotezę alternatywną.