

STATYSTYCZNE MODELOWANIE DANYCH BIOLOGICZNYCH

WYKŁAD 3



22 październik 2020

Plan wykładu

1. Test log-rank dla więcej niż dwóch grup
2. Test Mantela-Haenszela dla więcej niż dwóch grup
3. Wstęp do modelu proporcjonalnego hazardu Cox'a
4. Pakiet R

TEST LOG-RANK DLA WIĘCEJ NIŻ DWÓCH GRUP

Test log-rank dla więcej niż dwóch grup

Rozważmy 30 pacjentek chorych na raka szyjki macicy leczonych radioterapią (grupa A) i radioterapią z dodatkowym lekiem (grupa B). Dla tych pacjentek dodatkowo wprowadzamy zmienną określającą wiek pacjentki w momencie diagnozy. Następnie kobiety są dzielone na trzy grupy związane z wiekiem, w którym wykryto chorobę tzn. młode (Y) dla wieku mniejszego niż 50, w średnim wieku (M) dla przedziału wieku między 50, a 60 lat oraz seniorki (S) dla kobiet u których wykryto chorobę po 60 roku życia.

Test log-rank dla powyższego przykładu będzie miał postać:

$$\chi^2_{LR} = \frac{(O_Y - E_Y)^2}{E_Y} + \frac{(O_M - E_M)^2}{E_M} + \frac{(O_S - E_S)^2}{E_S}$$

W ogólnym przypadku test log-rank będzie miał postać:

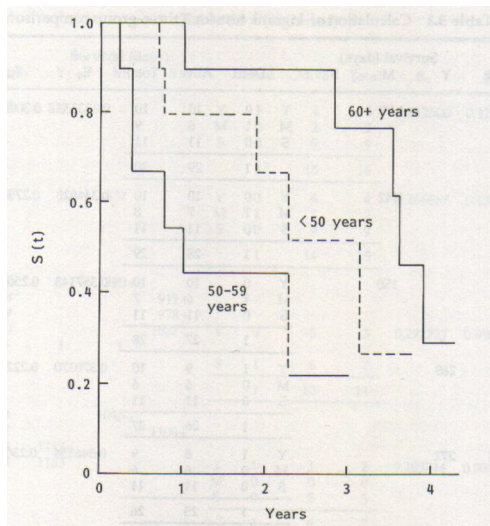
$$\chi^2_{LR} = \sum_{g=1}^G \frac{(O_g - E_g)^2}{E_g} \sim \chi^2_{G-1}$$

Test log-rank dla więcej niż dwóch grup

Control A				New Therapy B			
Patient number	Age (years)	Age group	Survival (days)	Patient number	Age (years)	Age group	Survival (days)
2	64	S	1037	1	63	S	1476+
4	69	S	1429	3	43	Y	827
5	38	Y	680	7	27	Y	519+
6	53	M	291	8	74	S	1100+
9	61	S	1577+	10	63	S	1307
11	59	M	90	12	39	Y	1360+
14	53	M	1090+	13	70	S	919+
15	56	M	142	16	68	S	373
17	65	S	1297	18	69	S	563+
19	48	Y	1113+	21	73	S	978+
20	45	Y	1153	22	58	M	650+
23	55	M	150	25	55	M	362
24	58	M	837	27	58	M	383+
26	49	Y	890+	28	45	Y	272
29	43	Y	269				
30	42	Y	468+				

Rysunek 1: 30 pacjentek chorych na raka szyjki macicy leczonych radioterapią (grupa A) i radioterapią z dodatkowym lekiem (grupa B) z podziałem na wiek, w którym wykryto chorobę.

Test log-rank dla więcej niż dwóch grup



Rysunek 2: Krzywe przeżycia Kaplana-Meiera dla 30 pacjentek chorych na raka szyjki macicy podzielonych ze względu na wiek, w którym wykryto chorobę.

Test log-rank dla więcej niż dwóch grup

Time t = 90	Dead	Alive	Total	$O_{Y,90}$	$O_{M,90}$	$O_{S,90}$	$E_{Y,90}$	$E_{M,90}$	$E_{S,90}$
Young	0	10	10	0	1	0	0.333333	0.300000	0.366667
Middle	1	8	9						
Senior	0	11	11						
Total	1	29	30						

Time t	Dead	Alive	Total	O_{Yt}	O_{Mt}	O_{St}	E_{Yt}	E_{Mt}	E_{St}
Young	a_t	d_t	m_t	a_t	b_t	c_t	$r_t m_t / N_t$	$r_t n_t / N_t$	$r_t l_t / N_t$
Middle	b_t	e_t	n_t						
Senior	c_t	f_t	l_t						
Total	r_t	s_t	N_t						

Rysunek 3: Przykład obliczeń obserwowanych i oczekiwanych śmierci w 90 dniu po wykryciu choroby dla 30 pacjentek chorych na raka szyjki macicy podzielonych ze względu na wiek, w którym wykryto chorobę.

Test log-rank dla więcej niż dwóch grup

Patient number	Rank	Survival (days)		Dead	Alive	Total	E_y	E_M	E_S
		Y	M						
11	1	90		Y 0	10	10	0.333333	0.300000	0.366667
				M 1	8	9			
				S 0	11	11			
				1	29	30			
15	2	142		Y 0	10	10	0.344828	0.275862	0.379310
				M 1	7	8			
				S 0	11	11			
				1	28	29			
23	3	150		Y 0	10	10	0.357143	0.250000	0.392857
				M 1	6	7			
				S 0	11	11			
				1	27	28			
29	4	269		Y 1	9	10	0.370370	0.222222	0.407607
				M 0	6	6			
				S 0	11	11			
				1	26	27			
28	5	272		Y 1	8	9	0.346154	0.230769	0.423077
				M 0	6	6			
				S 0	11	11			
				1	25	26			
6	6	291		Y 0	8	8	0.320000	0.240000	0.440000
				M 1	5	6			
				S 0	11	11			
				1	25	26			
25	7	362		Y 0	8	8	0.333333	0.208333	0.458333
				M 1	4	5			
				S 0	11	11			
				1	23	24			
16	8	373		Y 0	8	8	0.347826	0.173913	0.478261
				M 0	4	4			
				S 1	10	11			
				1	22	23			
27	9	383+							
30	10	468+							
7	11	519+							
18	12	563+							
22	13	650+							
5	14	680		Y 1	5	6	0.352941	0.117647	0.529412
				M 0	2	2			
				S 0	9	9			
				1	16	17			

Rysunek 4: Wyznaczanie testu log-rank dla 30 pacjentek chorych na raka szyjki macicy podzielonych ze względu na wiek, w którym wykryto chorobę.

Test log-rank dla więcej niż dwóch grup

Patient number	Rank	Survival (days)			Dead	Alive	Total	E_Y	E_M	E_S
		Y	M	S						
3	15	827			Y 1	4	5	0.312500	0.125000	0.562500
					M 0	2	2			
					S 0	9	9			
					1	15	16			
24	16	837			Y 0	4	4	0.266667	0.133333	0.600000
					M 1	1	2			
					S 0	9	9			
					1	14	15			
26	17	890+								
13	18		919+							
21	19		978+							
2	20		1037		Y 0	3	3	0.272727	0.090909	0.636364
M	0	1	1		S 1	6	7			
					1	10	11			
14	21		1090+							
8	22			1100+						
19	23	1113+			Y 1	1	2	0.285714	0.000000	0.714286
20	24	1153			M 0	0	0			
					S 0	5	5			
					1	6	7			
17	25		1297		Y 0	1	1	0.166667	0.000000	0.833333
					M 0	0	0			
					S 1	4	5			
					1	5	6			
10	26		1307		Y 0	1	1	0.200000	0.000000	0.800000
					M 0	0	0			
					S 1	3	4			
					1	4	5			
12	27	1360+			Y 0	0	0	0.000000	0.000000	1.000000
4	28		1429		M 0	0	0			
					S 1	2	3			
					1	2	3			
1	29		1476+							
9	30		1577+							
Totals					$E_Y = 4.610203$		$E_M = 2.367988$	$E_S = 9.021806$		
					$O_Y = 5$		$O_M = 6$	$O_S = 5$		
					$n_Y = 10$		$n_M = 9$	$n_S = 11$		

Rysunek 5: Wyznaczanie testu log-rank dla 30 pacjentek chorych na raka szyjki macicy podzielonych ze względu na wiek, w którym wykryto chorobę.

Test log-rank dla więcej niż dwóch grup

$$\chi_{LR}^2 = \frac{(5 - 4.610203)^2}{4.610203} + \frac{(6 - 2.367988)^2}{2.367988} + \frac{(5 - 9.021806)^2}{9.021806} = 7.40$$

Kwantyl rzędu 0.95 z rozkładu χ_2^2 wynosi 5.991465. Natomiast p-wartość wynosi 0.025 co sugeruje, że różnice pomiędzy krzywymi przeżycia dla powyższych trzech grup są statystycznie istotne.

TEST MANTELA-HAENSZELA DLA WIĘCEJ NIŻ DWÓCH GRUP

Test Mantela-Haenszela dla więcej niż dwóch grup

Statystyka Mantela-Haenszela dla dwóch grup może być przedstawiona w następujący sposób:

$$\chi_{MH}^2 = \frac{d^2}{V} = dV^{-1}d,$$

gdzie $d = (O_i - E_i)$ dla grupy A lub grupy B, a $V = \frac{\sum_t m_t n_t r_t s_t}{N_t(N_t - 1)^2}$.

Wystarczy wziąć d tylko dla jednej grupy ponieważ wyrażenia te są symetryczne i różnią się tylko znakiem tzn. $d_B = -d_A$. Jeśli wartości te podniesiemy do kwadratu to są one identyczne. Dodatkowo między wystąpieniami zgonów w grupie A (a_t) i B (b_t) występuje kowariancja, którą można wyznaczyć ze wzoru:

$$\text{Cov}(a_t, b_t) = -\frac{m_t n_t r_t s_t}{N_t(N_t - 1)^2}.$$

Minus w powyższym wzorze wynika z tego, że jeżeli zwiększymy a_t o jeden i jednocześnie zachowamy te same wartości brzegowe m_t , n_t , r_t i s_t to wartość b_t musi zmaleć o jeden.

Test Mantela-Haenszela dla więcej niż dwóch grup

Dla trzech grup ($G = 3$) będą potrzebne wariancje i kowariancje dla każdej pary a_t , b_t i c_t . Jednocześnie jeżeli zwiększymy a_t o jeden i zachowamy te same wartości brzegowe m_t , n_t , l_t , r_t i s_t to mamy wybór, którą wartość zmniejszyć b_t i c_t .

Używając oznaczeń z Rysunku (3) mamy:

$$Var(a_t) = \frac{m_t(n_t + l_t)r_t s_t}{N_t(N_t - 1)^2}$$

$$Var(b_t) = \frac{n_t(m_t + l_t)r_t s_t}{N_t(N_t - 1)^2}$$

$$Var(c_t) = \frac{l_t(n_t + m_t)r_t s_t}{N_t(N_t - 1)^2}$$

$$Cov(b_t, c_t) = -\frac{n_t l_t r_t s_t}{N_t(N_t - 1)^2}$$

Wartości $Cov(a_t, c_t)$ i $Cov(a_t, b_t)$ uzyskujemy z podobnych wzorów.

Test Mantela-Haenszela dla więcej niż dwóch grup

Całkowite wartości wariancji i kowariancji dla grup A, B i C wyznacza się poprzez sumę wariancji i kowariancji we wszystkich punktach czasu dla których wystąpiło zdarzenie.

Dla zwięzłości całkowite wariancje i kowariancje oznaczmy przez $Var(A)$, $Var(B)$, $Var(C)$, $Cov(A, B)$, $Cov(A, C)$ i $Cov(B, C)$. Test Mantela - Haenszela wymaga tylko dwóch grup więc dla wygody wybierzmy grupę A i B. Do opisu struktury kowariancji między tymi grupami możemy wtedy użyć postaci macierzowej:

$$\mathbf{V} = \begin{bmatrix} Var(A) & Cov(A, B) \\ Cov(B, A) & Var(B) \end{bmatrix}.$$

Test Mantela-Haenszela dla więcej niż dwóch grup

Teraz potrzebujemy macierzy dla korespondujących z macierzą V różnic $d_A = O_A - E_A$ i $d_B = O_B - E_B$:

$$\mathbf{d} = \begin{bmatrix} d_A \\ d_B \end{bmatrix}.$$

Ostatecznie statystyka Mantela-Haenszela ma postać:

$$\chi_{MH}^2 = \mathbf{d}^T \mathbf{V}^{-1} \mathbf{d}$$

Powyższy wzór można uogólnić na G grup. Wtedy $\mathbf{d}$ będzie miał $G - 1$ wierszy, natomiast macierz $\mathbf{V}$ będzie macierzą $(G - 1) \times (G - 1)$.

Test Mantela-Haenszela dla więcej niż dwóch grup

Niech $\mathbf{A}$ będzie macierzą kwadratową, wtedy macierz $\mathbf{B}$ jest macierzą odwrotną do $\mathbf{A}$ jeśli

$$\mathbf{AB} = \mathbf{BA} = \mathbf{I}$$

Dla macierzy nieosobliwej $\mathbf{A}$ mamy

$$\mathbf{A} = \frac{\mathbf{A}^D}{\det \mathbf{A}},$$

gdzie $\mathbf{A}^D$ jest transponowaną macierzą dopełnień algebraicznych do macierzy $\mathbf{A}$.

Test Mantela-Haenszela dla więcej niż dwóch grup

Macierz odwrotna do macierzy $\mathbf{V}$ ma postać

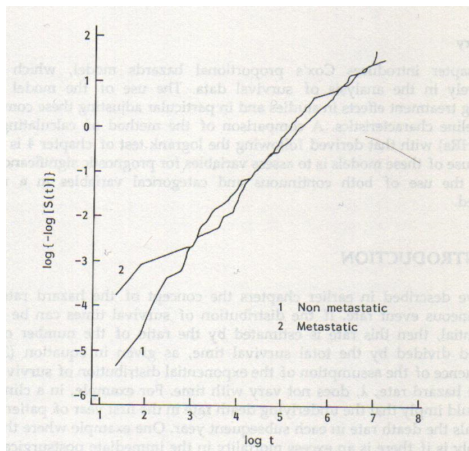
$$\mathbf{V}^{-1} = \begin{bmatrix} \text{Var}(B)/M & -\text{Cov}(A, B)/M \\ -\text{Cov}(B, A)/M & \text{Var}(A)/M \end{bmatrix},$$

gdzie $M = \text{Var}(A)\text{Var}(B) - [\text{Cov}(A, B)]^2$. I ostatecznie mamy:

$$\begin{aligned} \chi_{MH}^2 &= \mathbf{d}^T \mathbf{V}^{-1} \mathbf{d} \\ &= [d_A^2 \text{Var}(B) - d_A d_B \text{Cov}(A, B) - d_B d_A \text{Cov}(A, B) + d_B^2 \text{Var}(A)] / M \\ &= [d_A^2 \text{Var}(B) - 2d_A d_B \text{Cov}(A, B) + d_B^2 \text{Var}(A)] / M \end{aligned}$$

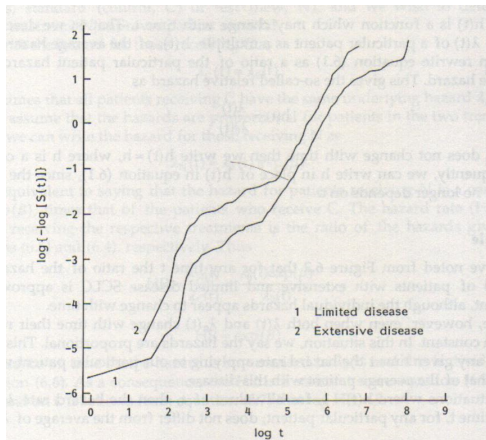
WSTĘP DO MODELU PROPORCJONALNEGO HAZARDU COX'A

Model Cox'a



Rysunek 6: Wykres $\log \{-\log [S_t]\}$ względem $\log t$ dla 304 pacjentów chorych na raka płuca (NSCLC) u których zdiagnozowano przerzuty bądź ich brak. Pacjenci byli leczeni radioterapią.

Model Cox'a



Rysunek 7: Wykres $\log \{-\log [S_t]\}$ względem $\log t$ dla pacjentów chorych na raka płuc (SCLC) z podziałem na pacjentów z rozległym i ograniczonym rakiem płuc. Pacjenci byli leczeni chemioterapią.

Podczas prowadzenia badań na grupie ludzi chorych na raka mózgu chcielibyśmy uzależnić długość ich przeżycia od wielu czynników m.in. wiek, przyjmowane lekarstwa lub stadium choroby. Aby to uczynić wygodne jest zdefiniowanie średniej funkcji hazardu dla pacjentów z rakiem mózgu $\lambda_0(t)$. Następnie możemy określić funkcję hazardu dla pojedynczego pacjenta z rakiem mózgu $\lambda(t)$. Dla funkcji $\lambda_0(t)$ i $\lambda(t)$ mamy zależność:

$$\lambda(t) = h(t)\lambda_0(t)$$

lub

$$h(t) = \frac{\lambda(t)}{\lambda_0(t)}.$$

Jeśli $h(t) = h$ to mówimy, że funkcja relatywnego hazardu $h(t)$ nie zależy od czasu t . Innymi słowy mówimy, że funkcje $\lambda_0(t)$ i $\lambda(t)$ są proporcjonalne.

Model proporcjonalnego hazardu Cox'a:

- często stosowana technika statystycznej analizy danych przeżycia;
- pozwala na wyizolowanie zmiennych objaśniających (niezależnych) mających wpływ na prognozę "failure time";
- można stosować nawet jeżeli zmienna zależna nie ma rozkładu normalnego oraz obserwacje są cenzurowane lub wykluczane.

Postać modelu:

$$\lambda(t, x_1(t), x_2(t), \dots, x_n(t)) = \lambda_0(t) \exp(\beta_1 x_1(t) + \dots + \beta_n x_n(t)),$$

gdzie

- $\lambda(t, x_1(t), x_2(t), \dots, x_n(t))$ - funkcja hazardu w momencie t przy n zmiennych objaśniających (zmienne objaśniające mogą być zarówno typu **ciągłego** jak i **dyskretnego**);
- $\lambda_0(t)$ - nieujemna funkcja hazardu odniesienia (**hazard bazowy**);
- funkcja $\lambda_0(t)$ nie zależy od żadnego parametru, jedynie od czasu t ;
- $\exp(\beta_1 x_1(t) + \dots + \beta_n x_n(t))$ - log-liniowa funkcja zależna jedynie od zmiennych objaśniających.

Z postaci funkcji hazardu wynikają następujące właściwości:

- funkcja hazardu przyjmuje wartości większe od 0;
- model Cox'a jest nieparametryczny, tzn. nie ma konieczności założenia z góry pewnego rozkładu zmiennej zależnej.

Zalety modelu Cox'a:

- **założenie proporcjonalności hazardu**: zakłada się, że dla dwóch obserwacji o różnych wartościach dla zmiennych objaśniających $x = (x_1, \dots, x_n)$ i $\tilde{x} = (\tilde{x}_1, \dots, \tilde{x}_n)$, stosunek funkcji hazardu dla tych dwóch obserwacji nie zależy bezpośrednio od czasu, a jedynie od zmiennych objaśniających, tzn.

$$\frac{\lambda(t, x_1(t), \dots, x_n(t))}{\lambda(t, \tilde{x}_1(t), \dots, \tilde{x}_n(t))} = \\ = \exp(\beta_1(x_1(t) - \tilde{x}_1(t)) + \dots + \beta_n(x_n(t) - \tilde{x}_n(t)))$$

Oznacza to, że jeżeli przypatrujemy się dwóm niezwiązanym ze sobą obiektom w tym samym czasie, to różnice w natężeniu zdarzeń są zależne od różnic między zmiennymi objaśniającymi.

- znając współczynniki β_j jesteśmy w stanie, bez jakiejkolwiek wiedzy o hazardzie bazowym, określić wrażliwość funkcji hazardu na zmiany j -tej cechy. Mianowicie, gdy j -ta zmienna wzrośnie o jednostkę (a pozostałe będą bez zmian), to funkcja hazardu wzrośnie o $\exp(\beta_j)$ razy.
- taka postać $\lambda(t)$ pozwala na estymację współczynników β_j przy minimalnych założeniach dla hazardu bazowego.

Zakładając, że $\lambda_0(t)$ jest dowolne nie możemy dostać żadnych informacji o współczynnikach β_j na podstawie badań okresów, w których nie za-
szło żadne zdarzenie, ponieważ w tych przedziałach funkcja $\lambda_0(t)$ może
być równa 0. Dlatego musimy rozważać prawdopodobieństwa warunkowe:
 $P(T = t_{(i)} | \mathcal{R}(t_{(i)}))$ wystąpienia zdarzenia w momencie $t_{(i)}$ na zbiorze
 $\mathcal{R}(t_{(i)})$ wszystkich takich obserwacji, dla których “failure time” lub “czas
ocenzurowania” jest co najmniej równy $t_{(i)}$.

Funkcja wiarygodności:

$$\begin{aligned} P(T = t_{(i)} | \mathcal{R}(t_{(i)})) &= \frac{P(T = t_{(i)} | T \geq t_{(i)})}{\sum_{j \in \mathcal{R}_i} P(T = t_{(j)} | T \geq t_{(j)})} = \\ &= \frac{\lambda(t_{(i)}, x_{(i)})}{\sum_{j \in \mathcal{R}_i} \lambda(t_{(j)}, x_{(j)})} = \frac{\exp(\beta x_{(i)})}{\sum_{j \in \mathcal{R}_i} \exp(\beta x_{(j)})} = \\ &= \frac{\exp(\beta_1 x_{(i)1} + \dots + \beta_n x_{(i)n})}{\sum_{j \in \mathcal{R}_i} \exp(\beta_1 x_{(j)1} + \dots + \beta_n x_{(j)n})}, \end{aligned}$$

gdzie $x_{(j)}$ to j -ta obserwacja, której "failure time" przypadł na czas $t_{(i)}$.

Znajdując maksimum funkcji log-wiarygodności można wyznaczyć estymator współczynników β_j .

$$L(\beta) = \sum_{i=1}^k \beta x_{(i)} - \sum_{i=1}^k \log \left(\sum_{j \in \mathcal{R}_i} \exp(\beta x_{(j)}) \right)$$

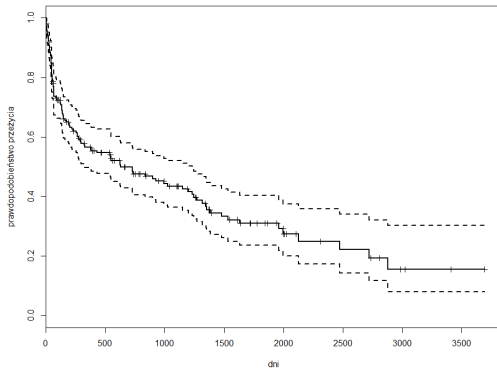
gdzie $i = 1, \dots, k$ to wszystkie momenty wystąpienia zdarzenia.

PAKIET R



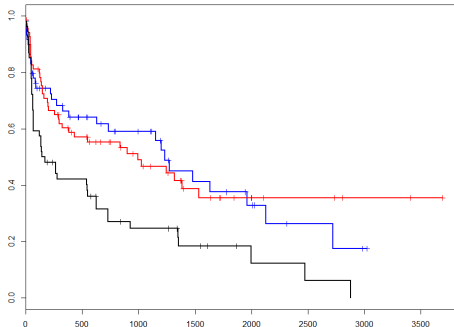
Pakiet R

```
1 library(survival)
2 data(stanford2)
3 attach(stanford2)
4 plot(survfit(Surv(time,status)~1),xlab="dni",lwd=2)
```



Pakiet R

```
1 q1=quantile(age,0.33)
2 q2=quantile(age,0.66)
3 age2=rep(2,184)
4 age2[age<q1]=1
5 age2[age>q2]=3
6 model=survfit(Surv(time,status)~age2)
7 plot(model)
8 plot(model,lwd=2,col=c("blue","red","black"))
```



```
1 survdiff(Surv(time,status)~age2)
```

Call:

```
survdiff(formula = Surv(time, status) ~ age2)
```

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
age2=1	60	31	38.3	1.38	2.10
age2=2	70	38	46.6	1.60	2.74
age2=3	54	44	28.1	9.02	12.13

Chisq= 12.1 on 2 degrees of freedom, p= 0.00233

```
1 model=coxph(Surv(time,status)~age+t5)
2 summary(model)
```

```
Call:
coxph(formula = Surv(time, status) ~ age + t5)

n= 157, number of events= 102
(27 observations deleted due to missingness)

      coef exp(coef) se(coef)      z Pr(>|z|)
age 0.02961  1.03006  0.01136  2.608  0.00911 **
t5  0.17041  1.18579  0.18326  0.930  0.35243
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

      exp(coef) exp(-coef) lower .95 upper .95
age      1.030      0.9708      1.007      1.053
t5      1.186      0.8433      0.828      1.698

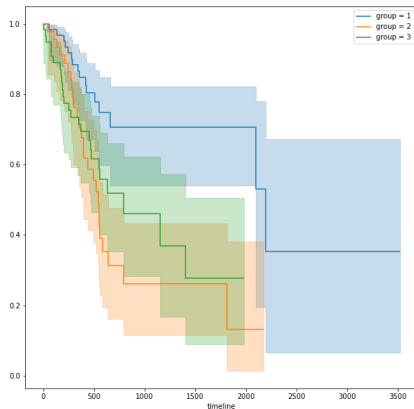
Concordance= 0.59 (se = 0.031 )
Rsquare= 0.053 (max possible= 0.997 )
Likelihood ratio test= 8.47 on 2 df,  p=0.01449
Wald test               = 7.81 on 2 df,  p=0.02018
Score (logrank) test = 7.87 on 2 df,  p=0.0195
```



```
from lifelines import *
import kaplanmeier as km
from lifelines.plotting import add_at_risk_counts
import matplotlib.pyplot as plt

## example data ##
df = km.example_data()
time = df['time']
event = df['Died']
groups = df['group']
groups[:3] = 3
i1 = (groups == 1)
i2 = (groups == 2)
i3 = (groups == 3)
print(groups)

plt.figure(figsize = (10, 10))
kmf = KaplanMeierFitter()
kmf.fit(time[i1], event_observed = event[i1], label = 'group = 1')
kmf.plot()
kmf.fit(time[i2], event_observed = event[i2], label = 'group = 2')
kmf.plot()
kmf.fit(time[i3], event_observed = event[i3], label = 'group = 3')
kmf.plot()
plt.show()
```



Rysunek 8: Trzy krzywe przeżycia na jednym wykresie w Pythonie.

```
cox = CoxPHFitter()  
cox.fit(df, 'time', event_col = 'Died')  
cox.print_summary()
```

Rysunek 9: Model Cox'a w Pythonie.

model	lifelines.CoxPHFitter										
duration col	'time'										
event col	'Died'										
baseline estimation	breslow										
number of observations	180										
number of events observed	67										
partial log-likelihood	-291.76										
time fit was run	2020-10-21 13:47:17 UTC										
	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p	-log2(p)	
group	0.38	1.47	0.15	0.10	0.67	1.10	1.95	2.62	0.01	6.85	
Concordance	0.60										
Partial AIC	585.52										
log-likelihood ratio test	7.02 on 1 df										
-log2(p) of ll-ratio test	6.96										

Rysunek 10: Wyniki w modelu Cox'a w Pythonie.

Wykład opracowany na podstawie książki:

- Mahesh K. B. Parmer and David Machin “Survival Analysis - A Practical Approach”