Temat projektu: Modele parametryczne w analizie historii zdarzeń

Autor: Sebastian Boruch

Spis treści

1. Opis danych

Australijskie badanie przeprowadzone przez Caplehorna i Bell (1991) z departamentu zdrowia publicznego Uniwersytetu w Sydney porównywało czas przeżycia w dwóch klinikach leczenia metadonem dla osób uzależnionych od heroiny. Czas przeżycia pacjenta określono jako czas w dniach, aż pacjent powrócił do nałogu lub opuścił klinikę. Obie kliniki różniły się w zależności od ogólnej polityki leczenia. Celem było zidentyfikowanie czynników, które wpływają na czas przeżycia: klinika, maksymalna dzienna dawka metadonu i (nie)obecność w więzieniu.

Proces leczenia pacjentów metadonem badano w kohorcie 238 osób uzależnionych od heroiny, którzy weszli do programu terapii uzależnień od lutego 1986 do sierpnia 1987 r. Wszyscy pacjenci zostali ocenieni w tym samym instytucie i odesłani do jednej z dwóch placówek leczniczych w celu rozpoczęcia terapii.

Rozkład Weibulla to ciągły rozkład prawdopodobieństwa często stosowany w analizie przeżycia do modelowania sytuacji, gdy prawdopodobieństwo awarii zmienia się w czasie.

Może on w zależności od parametrów przypominać zarówno rozkład normalny, jak i rozkład wykładniczy (sprowadza się do niego dla k=1). Model wykładniczy jest szczególnym przypadkiem modelu Weibulla. Model przedziałami stały zakłada, że funkcja hazardu jest przedziałami stała.

W programie SAS rozkład Weibulla wykorzystywany jest za pomocą procedury LIFEREG. Procedura LIFEREG umożliwia parametryczną estymację czasu porażki. Analizowane dane mogą być cenzurowane prawostronnie, lewostronnie bądź przedziałowo. Procedura ta wykorzystuje algorytm Newtona-Raphsona w celu estymacji parametrów, używając maksymalnej wiarygodności. Procedura uniemożliwia uwzględnienie zmiennych zależnych od czasu. Obserwacje nie mogą zawierać braków danych dla zmiennej zależnej – obserwacja zostaje wtedy ocenzurowana. Usuwane są również obserwacje które mają braki danych przy zmiennych objaśniających.

1.1. Kategoryzacja zmiennej DAWKA

Trudności w interpretacji może sprawiać pierwotna postać zmiennej DAWKA- jest to zmienna ilościowa z 15 poziomami. Dla potrzeb tej analizy, w celu ułatwienia interpretacji, zmienna DAWKA została skategoryzowana na 3 kategorie:

- poniżej 60 mg/dzień
- 60 mg/dzień
- powyżej 60/ dzień

Jest to kategoryzacja zgodna z rozkładem tej zmiennej- dominantą w rozkładzie była wartość 60, a pozostałe dwie kategorie są równoliczne. Rozkład zmiennej był zbliżony do normalnego.

2. Interpretacja wyników procedury LIFEREG

2.1. Wyniki modelu z rozkładem wykładniczym bez zmiennych

model z rozkładem wykładniczym bez zmiennych

Procedura LIFEREG

Informacje o modelu	
Zbiór	WORK.ADDICTS
Zmienna zależna	Log(Dni przeżycia)
Zmienna obcięcia	Status
Wartości obcięcia	0
Liczba obserwacji	238
Wartości nieobcięte	150
Wartości obcięte prawostronnie	88
Wartości obcięte lewostronnie	0
Wartości obcięte w przedziale	0
Liczba parametrów	1
Name of Distribution	Exponential
Log. wiarygodności	-295.4343382

Wczytano obserwacji	238
Użyto obserwacji	238

Statystyki dopasowania	
-2 log. wiarygodności	590.869
AIC (jak najmniejsze)	592.869
AICC (jak najmniejsze)	592.886
BIC (jak najmniejsze)	596.341

Statystyki dopasowania (odpowiedź nierejestrowana)	
-2 log. wiarygodności	2237.852
ExponentialAIC (jak najmniejsze)	2239.852
ExponentialAICC (jak najmniejsze)	2239.869
ExponentialBIC (jak najmniejsze)	2243.325

Algorytm osiągnął zbieżność.

Tabela 1. Podstawowe statystyki modelu z rozkładem wykładniczym bez zmiennych

Występuje tu tylko jeden parametr – parametr α. Logarytm funkcji wiarygodności nie jest w tym modelu wprost interpretowalny, ale będzie wykorzystany do późniejszego porównania modeli między sobą (TSW). Warto dodać, że każdy model dąży do maksymalizacji funkcji wiarygodności, więc model z większym log. wiarygodności uznaje się za lepszy.

Statystyki dopasowania- najpierw widoczne są wyniki dla logarytmu zmiennej zależnej (log zmiennej czasowej), a następnie dla niezlogarytmowanej zmiennej czasowej. Wyniki te w tym momencie nie są interpretowalne. Będą one porównywane pomiędzy modelami. Wybiera się ten model, który ma jak najmniejsze wartości statystyk dopasowania.

Analiza ocen parametrów maksymalnej wiarygodności							
Parametr	DF	Ocena	Błąd standardowy	Przedział ufności 95%		Chi-kwadrat	Pr. > chi-kw.
Intercept	1	6.4595	0.0816	6.2995	6.6195	6258.79	<.0001
Skala	0	1.0000	0.0000	1.0000	1.0000		
Skala Weibulla	1	638.7466	52.1534	544.2874	749.5988		
Kształt Weibulla	0	1.0000	0.0000	1.0000	1.0000		

Statystyki mnożnika Lagrange'a		
Parametr	Chi-kwadrat	Pr. > chi-kw.
Skala	11.4745	0.0007

Tabela 2. Analiza ocen parametrów modelu z rozkładem wykładniczym bez zmiennych

Parametry zostały oszacowane za pomocą metody największej wiarygodności. Intercept i Skala to parametry rozkładu zmiennych ekstremalnych. Aby na ich podstawie uzyskać wartości parametrów α i β, należy przeprowadzić następujące obliczenia:

- α = exp(-Intercept) → α = exp(-6,4595) = 0.0015
- β = 1/Skala → β = 1/1 = 1

Parametry funkcji gęstości, przeżycia i hazardu:

- $\hat{f}(t) = 0{,}0015 * exp(-0{,}0015t)$
- $\hat{S}(t) = exp(-0{,}0015t)$
- $\hat{h}(t) = 0{,}0015$

Skala Weibulla to inaczej exp(Intercept)- interpretuje się tę wartość jako średni czas do zajścia zdarzenia (powrotu pacjenta do nałogu), który wynosi 638,75 dni.

Badana jest również istotność wyrazu wolnego (Intercept):

- H0 - nieistotny statystycznie
- H1 - istotny statystycznie

Na poziomie istotności 5% można odrzucić H0 na rzecz hipotezy alternatywnej, która uznaje parametr za istotny statystycznie.

Test mnożników Lagrange'a - jest to test punktowy, który bada hipotezę, że parametr Skala (β) = 1. Zatem, test bada czy model wykładniczy jest dobrym modelem do dopasowania do empirycznego rozkładu zmiennej czasowej. Test ten weryfikuje, czy parametr powinien być równy 1.

- H0: β = 1
- H1: β ≠ 1

Na poziomie istotności 5% można odrzucić H0 na rzecz hipotezy alternatywnej, zatem można stwierdzić że rozkład wykładniczy nie jest odpowiednim rozkładem z punktu widzenia dopasowania do zmiennej czasowej.

2.2. Wyniki modelu z rozkładem wykładniczym ze zmiennymi objaśniającymi

Procedura LIFEREG

Informacje o modelu	
Zbiór	WORK.QUERY_FOR_ADDICTS
Zmienna zależna	Log(Dni przeżycia)
Zmienna obcięcia	Status
Wartości obcięcia	0
Liczba obserwacji	238
Wartości nieobcięte	150
Wartości obcięte prawostronnie	88
Wartości obcięte lewostronnie	0
Wartości obcięte w przedziale	0
Liczba parametrów	5
Name of Distribution	Exponential
Log. wiarygodności	-272.989263

Wczytano obserwacji	238
Użyto obserwacji	238

Informacje o poziomach klasyfikacji		
Nazwa	Poziomy	Wartości
Dawka	3	60 <60 >60
Klinika	2	1 2
Więzienie	2	0 1

Statystyki dopasowania	
-2 log. wiarygodności	545.979
AIC (jak najmniejsze)	555.979
AICC (jak najmniejsze)	556.237
BIC (jak najmniejsze)	573.340

Statystyki dopasowania (odpowiedź nierejestrowana)	
-2 log. wiarygodności	2192.962
ExponentialAIC (jak najmniejsze)	2202.962
ExponentialAICC (jak najmniejsze)	2203.221
ExponentialBIC (jak najmniejsze)	2220.324

Tabela 3. Podstawowe statystyki modelu z rozkładem wykładniczym ze zmiennymi

Powyższe statystyki jak i sposób interpretacji są analogiczne dla modelu bez zmiennych. Statystyki dopasowania zostaną porównane w dalszej części analizy. Dodatkową informacją jest tu informacja o poziomach klasyfikacji. Zmienna DAWKA posiada 3 poziomy po kategoryzacji. Zmienne KLINIKA i WIĘZIENIE to zmienne binarne. Poziom 1 dla KLINIKA oznacza odbycie przez pacjenta terapii w klinice nr 1, poziom 2- w klinice nr 2. Poziom 0 dla WIĘZIENIE oznacza, że pacjent nigdy nie przebywał w więzieniu, poziom 1- że przebywał.

Analiza efektów typu III			
Efekt	DF	Chi-kwadrat Walda	Pr. > chi-kw.
Dawka	2	16.7009	0.0002
Klinika	1	17.8638	<.0001
Więzienie	1	2.1440	0.1431

Analiza ocen parametrów maksymalnej wiarygodności								
Parametr		DF	Ocena	Błąd standardowy	Przedział ufności 95%		Chi-kwadrat	Pr. > chi-kw.
Intercept		1	7.3623	0.2261	6.9191	7.8054	1060.33	<.0001
Dawka	60	1	-0.4508	0.2235	-0.8888	-0.0128	4.07	0.0437
Dawka	<60	1	-0.7817	0.1913	-1.1566	-0.4068	16.70	<.0001
Dawka	>60	0	0.0000	.	.	.	.	.
Klinika	1	1	-0.9026	0.2135	-1.3211	-0.4840	17.86	<.0001
Klinika	2	0	0.0000	.	.	.	.	.
Więzienie	0	1	0.2437	0.1664	-0.0825	0.5699	2.14	0.1431
Więzienie	1	0	0.0000	.	.	.	.	.
Skala		0	1.0000	0.0000	1.0000	1.0000		
Kształt Weibulla		0	1.0000	0.0000	1.0000	1.0000		

Statystyki mnożnika Lagrange'a		
Parametr	Chi-kwadrat	Pr. > chi-kw.
Skala	32.7829	<.0001

Tabela 4. Analiza parametrów i efektów modelu z rozkładem wykładniczym ze zmiennymi

Analiza efektów typu III- występują tu 3 zmienne: DAWKA (2 stopnie swobody) oraz KLINIKA i WIĘZIENIE (po jednym stopniu swobody). Liczba kategorii to stopnie swobody + 1. Na podstawie tabeli efektów bada się istotność zmiennych:

- H0 – nieistotna statystycznie
- H1 –istotna statystycznie.

Na poziomie istotności 5% odrzuca się H0 na rzecz H1. Zatem zmienne KLINIKA i DAWKA są istotne statystycznie, natomiast zmienna WIĘZIENIE jest nieistotna.

Analiza ocen parametrów- do modelu zostały włączone 3 zmienne, ale jedna z nich- WIĘZIENIE jest nieistotna statystycznie, więc nie ma powodu aby interpretować jej wynik.

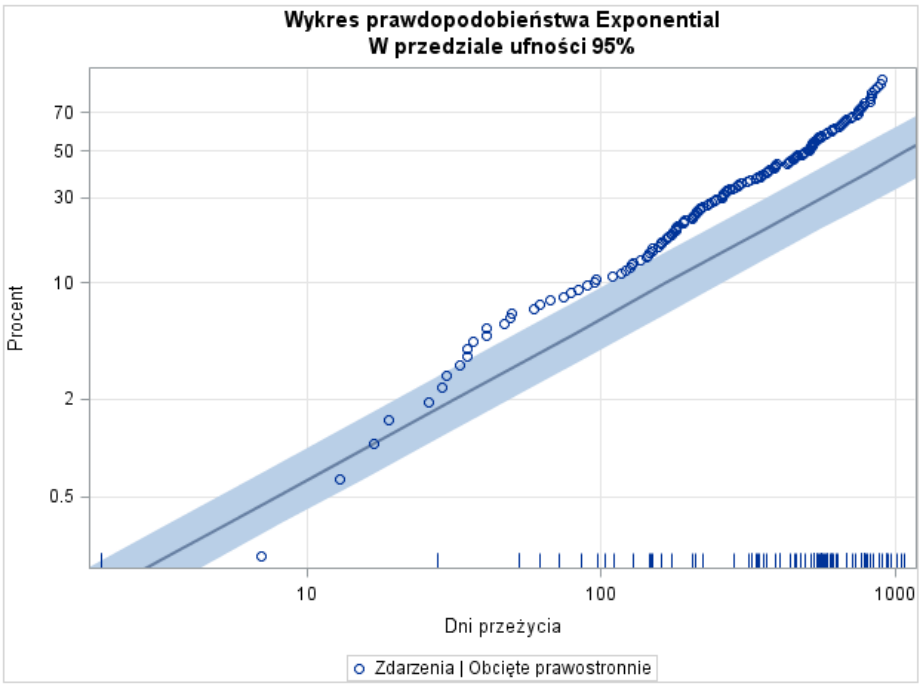
Interpretacja ocen zmiennej DAWKA: jest to zmienna klasyfikująca z trzema kategoriami. Wartością referencyjną jest tu „powyżej 60 mg/dzień metadonu”. Interpretacja wyników jest następująca:

- przeciętny czas powrotu do nałogu wśród pacjentów przyjmujących dawkę poniżej 60 mg/dzień jest 100%- exp(-0.7817)*100%= 54% krótszy niż w przypadku tych przyjmujących powyżej 60 mg/dzień
- ryzyko powrotu do nałogu u przyjmujących dawkę poniżej 60 mg/dzień jest o exp(-(-0.7818)-100% = 118% wyższe niż u osób przyjmujących powyżej 60 mg/dzień
- przeciętny czas powrotu do nałogu wśród pacjentów przyjmujących dawkę 60 mg/dzień jest 100%- exp(-0.4508)*100%= 36% krótszy niż w przypadku tych przyjmujących powyżej 60 mg/dzień
- ryzyko powrotu do nałogu u przyjmujących dawkę 60 mg/dzień jest o exp(-(-0.4508)-100% = 57%% wyższe niż u osób przyjmujących powyżej 60 mg/dzień

Interpretacja wyników zmiennej KLINIKA (wartość referencyjna- klinika 2):

- przeciętny czas powrotu do nałogu u pacjentów z kliniki 1 jest 100%-exp(-0.9026)=59% krótszy niż u pacjentów z kliniki 2

- ryzyko powrotu do nałogu u pacjentów z kliniki 1 jest o exp(-(-0.9026)-100%=147% wyższe niż u pacjentów z kliniki 2



Wykres 1. Wykres prawdopodobieństwa modelu z rozkładem wykładniczym ze zmiennymi

2.3. Wyniki modelu z rozkładem Weibulla

Wyniki modelu z rozkładem Weibulla są podobne do tych z modelu z rozkładem wykładniczym. Występują tu takie same zmienne objaśniające z taką samą liczbą stopni swobody. Inna jest wartość parametru Skala i oceny parametrów nieco się różnią od modelu wykładniczego. Interpretacja wyników informacji o modelu i statystyk dopasowania została opisana w poprzednich modelach. Wartości statystyk zostaną użyte do porównania w teście stosunku wiarygodności w dalszej części analizy.

Informacje o modelu	
Zbiór	WORK.QUERY_FOR_ADDICTS
Zmienna zależna	Log(Dni przeżycia)
Zmienna obcięcia	Status
Wartości obcięcia	0
Liczba obserwacji	238
Wartości nieobcięte	150
Wartości obcięte prawostronnie	88
Wartości obcięte lewostronnie	0
Wartości obcięte w przedziale	0
Liczba parametrów	6
Name of Distribution	Weibull
Log. wiarygodności	-264.1664509

Wczytano obserwacji	238
Użyto obserwacji	238

Informacje o poziomach klasyfikacji		
Nazwa	Poziomy	Wartości
Dawka	3	60 <60 >60
Klinika	2	1 2
Więzienie	2	0 1

Statystyki dopasowania	
-2 log. wiarygodności	528.333
AIC (jak najmniejsze)	540.333
AICC (jak najmniejsze)	540.697
BIC (jak najmniejsze)	561.167

Statystyki dopasowania (odpowiedź nierejestrowana)	
-2 log. wiarygodności	2175.317
WeibullAIC (jak najmniejsze)	2187.317
WeibullAICC (jak najmniejsze)	2187.680
WeibullBIC (jak najmniejsze)	2208.150

Tabela 5. Podstawowe statystyki modelu z rozkładem Weibulla ze zmiennymi

Analiza efektów typu III			
Efekt	DF	Chi-kwadrat Walda	Pr. > chi-kw.
Dawka	2	21.8571	<.0001
Klinika	1	20.8543	<.0001
Więzienie	1	3.6191	0.0571

Analiza ocen parametrów maksymalnej wiarygodności								
Parametr		DF	Ocena	Błąd standardowy	Przedział ufności 95%		Chi-kwadrat	Pr. > chi-kw.
Intercept		1	7.1227	0.1710	6.7875	7.4579	1734.22	<.0001
Dawka	60	1	-0.3474	0.1659	-0.6726	-0.0223	4.39	0.0363
Dawka	<60	1	-0.6670	0.1430	-0.9473	-0.3867	21.76	<.0001
Dawka	>60	0	0.0000	.	.	.	.	.
Klinika	1	1	-0.7357	0.1611	-1.0515	-0.4200	20.85	<.0001
Klinika	2	0	0.0000	.	.	.	.	.
Więzienie	0	1	0.2338	0.1229	-0.0071	0.4747	3.62	0.0571
Więzienie	1	0	0.0000	.	.	.	.	.
Skala		1	0.7374	0.0501	0.6454	0.8425		
Kształt Weibulla		1	1.3561	0.0922	1.1869	1.5494		

Tabela 6. Analiza parametrów i efektów modelu z rozkładem Weibulla ze zmiennymi

Zmienna WIĘZIENIE ponownie nie jest istotna statystycznie, ale tym razem balansuje na granicy istotności. Zmienne DAWKA i KLINIKA są istotne statystycznie na poziomie 5%.

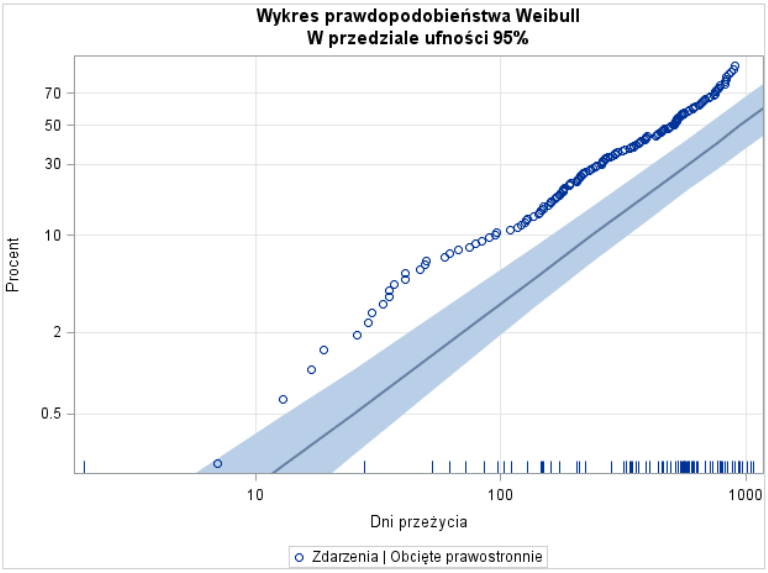
Interpretacja ocen zmiennej DAWKA: wartością referencyjną jest tu „powyżej 60 mg/dzień metadonu". Interpretacja wyników jest następująca:

- przeciętny czas powrotu do nałogu wśród pacjentów przyjmujących dawkę poniżej 60 mg/dzień jest 100%- exp(-0.6670)*100%= 49% krótszy niż w przypadku tych przyjmujących powyżej 60 mg/dzień
- ryzyko powrotu do nałogu u przyjmujących dawkę poniżej 60 mg/dzień jest o exp(-(-0.6670)-100% = 95% wyższe niż u osób przyjmujących powyżej 60 mg/dzień
- przeciętny czas powrotu do nałogu wśród pacjentów przyjmujących dawkę 60 mg/dzień jest 100%- exp(-0.3474)*100%= 29% krótszy niż w przypadku tych przyjmujących powyżej 60 mg/dzień
- ryzyko powrotu do nałogu u przyjmujących dawkę 60 mg/dzień jest o exp(-(-0.4508)-100% = 41% wyższe niż u osób przyjmujących powyżej 60 mg/dzień

Interpretacja wyników zmiennej KLINIKA (wartość referencyjna- klinika 2):

- przeciętny czas powrotu do nałogu u pacjentów z kliniki 1 jest 100%-exp(-0.7357)=52% krótszy niż u pacjentów z kliniki 2
- ryzyko powrotu do nałogu u pacjentów z kliniki 1 jest o exp(-(-0.7357)-100%=109% wyższe niż u pacjentów z kliniki 2

Wartość parametru $\hat{\beta} = \frac{1}{Skala} = \frac{1}{0{,}7374} = 1{,}35$ jest większa od 1, a więc funkcja hazardu jest rosnąca.



Wykres 2. Wykres prawdopodobieństwa dla modelu Weibulla ze zmiennymi

2.4. Test stosunku wiarygodności (TSW)

Korzystając z testu stosunku wiarygodności autor dokona teraz porównania modelu wykładniczego ze zmiennymi objaśniającymi i modelu Weibulla ze zmiennymi.

- Model 1- model wykładniczy ze zmiennymi (liczba parametrów: 5); - log likelihood= -273
- Model 2- model Weibulla ze zmiennymi (liczba parametrów: 6); - log likelihood= -264

Hipotezy:

- H0: Model 1 jest lepszy niż Model 2 – parametry w 2 modelu przy zmiennej są równe 0
- H1: Model 2 jest lepszy niż model

TSW=−2*($LogLikelihood(model\ 1)$− $LogLikelihood(model\ 2)$)
TSW= -2*(-272,9892 + 264,1164) = 17,74

W tym porównaniu liczba stopni swobody wynosi: 6-2 = 4.
Na poziomie istotności α=0,05 wartość krytyczna testu wynosi $X^2_{0,05;4}$=9,488. Wartość 17,74 wpada w obszar odrzuceń. Odrzuca się H0 na korzyść H1. Model 2 jest lepszy niż Model 1.

3. Wnioski

Najlepiej dopasowanym okazał się model z rozkładem Weibulla ze zmiennymi objaśniającymi. Oszacował on, że dawka powyżej 60 mg/dzień jest najkorzystniejsza dla pacjentów walczących z narkomanią, a pobyt w klinice 2 jest korzystniejszy niż w klinice 1 oraz, że pobyt w więzieniu na poziomie istotności 5% nie ma wpływu na powrót do nałogu. Model oszacował, że funkcja hazardu rośnie, więc ryzyko powrotu do nałogu rośnie w czasem.

4. Kod SAS

```
PROC FORMAT
      LIB=WORK;
            VALUE dawka
                  0 - 59 = "<60"
                  60 = "60"
                  61 - 120 = ">60";
RUN
;
PROC SQL;
   CREATE TABLE WORK.QUERY_FOR_ADDICTS AS
   SELECT t1.Dawka FORMAT=DAWKA. AS Dawka,
          t1.Klinika,
          t1.Status,
          t1.'Więzienie'n,
          t1.'Dni przeżycia'n
      FROM WORK.ADDICTS t1;
QUIT;

ODS GRAPHICS ON;
title1 "model z rozkładem wykładniczym bez zmiennych";
PROC LIFEREG data = WORK.ADDICTS;
MODEL "Dni przeżycia"n*STATUS (0) = / dist=exponential;
RUN;
title;

title1 "model z rozkładem wykładniczym ze zmiennymi";
PROC LIFEREG data = WORK.QUERY_FOR_ADDICTS;
CLASS DAWKA Klinika 'Więzienie'n;
MODEL "Dni przeżycia"n*STATUS (0) = DAWKA klinika 'Więzienie'n/
dist=exponential;
probplot;
RUN;
```

```
TITLE1 "model Weibulla ze zmiennymi";
proc lifereg data=WORK.QUERY_FOR_ADDICTS;
class DAWKA Klinika 'Więzienie'n;
model "Dni przeżycia"n*STATUS (0)=DAWKA Klinika
'Więzienie'n/dist=weibull;
probplot;
run;
title;
ods graphics off;
```