Komisja Egzaminacyjna dla Aktuariuszy

LXXXI Egzamin dla Aktuariuszy

Sesja egzaminacyjna w dniu 19 listopada 2019 r.

Modelowanie

Imię i nazwisko osoby egzaminowanej: .....

Czas trwania egzaminu: 120 minut

Uwagi

a) W prezentowanych wynikach separatorem dziesiętnym (znakiem dziesiętnym) jest kropka „ .”.

b) W prezentowanych wynikach oszacowań uogólnionych modeli liniowych (GLM):
- Residual deviance i Resid. Dev – oznacza dewiancję oszacowanego modelu,
- Null deviance – oznacza dewiancję modelu zerowego,
- Deviance – redukcję dewiancji po dodaniu kolejnej zmiennej objaśniającej,
- Df – stopnie swobody,
- Sum Sq – suma kwadratów.

c) W zadaniach wartość zagrożona na poziomie ufności α jest definiowana jako kwantyl rzędu α rozkładu odpowiedniej zmiennej losowej, tzn.

$$VaR_{\alpha}(X) = \inf\{x: F_X(x) \geq \alpha\}\,.$$

d) W zadaniach zastosowano następujące oznaczenia:
$E(X)$ – wartość oczekiwana
$D(X)$ – odchylenie standardowe

e) Wartości $\chi^2_{\alpha;\upsilon}$ rozkładu chi-kwadrat spełniające warunek $P(\chi^2 \geq \chi^2_{\alpha;\upsilon}) = \alpha$

υ \ α	***0.995***	***0.99***	***0.975***	***0.95***	***0.9***	***0.1***	***0.05***	***0.025***	***0.01***	***0.005***
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801

f) Wybrane kwantyle standardowego rozkładu normalnego

α	***0.950***	***0.980***	***0.981***	***0.982***	***0.983***	***0.984***	***0.985***	***0.986***	***0.987***	***0.988***
u_α	1.645	2.054	2.075	2.097	2.120	2.144	2.170	2.197	2.226	2.257

α	***0.989***	***0.990***	***0.991***	***0.992***	***0.993***	***0.994***	***0.995***	***0.996***	***0.997***	***0.998***
u_α	2.290	2.326	2.366	2.409	2.457	2.512	2.576	2.652	2.748	2.878

g) Wybrane kwantyle rozkładu gamma z parametrami $\alpha = 2$ i $\beta = 50$ (przyjęto parametryzację, w której $E(X) = \alpha\beta$, $D^2(X) = \alpha\beta^2$)

p	***0,980***	***0,981***	***0,982***	***0,983***	***0,984***	***0,985***	***0,986***	***0,987***	***0,988***	***0,989***
$F^{-1}(p)$	291,70	294,70	297,86	301,19	304,72	308,48	312,48	316,78	321,41	326,43

p	***0,990***	***0,991***	***0,992***	***0,993***	***0,994***	***0,995***	***0,996***	***0,997***	***0,998***	***0,999***
$F^{-1}(p)$	331,92	337,97	344,72	352,36	361,15	371,51	384,14	400,36	423,09	461,67

h) Wartość oczekiwana i wariancja rozkładu logarytmiczno-normalnego o parametrach μ i σ:

$$E(X) = e^{\mu+\frac{1}{2}\sigma^2}$$

$$D^2(X) = \left(e^{\sigma^2} - 1\right)e^{2\mu+\sigma^2}$$

i) Wartości F_{r_1,r_2} rozkładu F spełniające warunek $P(F \geq F_{r_1,r_2}) = 0.05$

r_2 \ r_1	***1***	***2***	***3***	***4***	***5***	***6***	***7***
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103
700	3.855	3.009	2.618	2.385	2.227	2.112	2.023
800	3.853	3.007	2.616	2.383	2.225	2.110	2.021
900	3.852	3.006	2.615	2.382	2.224	2.109	2.020
1000	3.851	3.005	2.614	2.381	2.223	2.108	2.019
∞	3.844	2.998	2.607	2.374	2.216	2.101	2.012

r_2 \ r_1	***8***	***9***	***10***	***11***	***12***	***13***	***14***
1	238.883	240.543	241.882	242.983	243.906	244.690	245.364
2	19.371	19.385	19.396	19.405	19.413	19.419	19.424
3	8.845	8.812	8.786	8.763	8.745	8.729	8.715
4	6.041	5.999	5.964	5.936	5.912	5.891	5.873
5	4.818	4.772	4.735	4.704	4.678	4.655	4.636
6	4.147	4.099	4.060	4.027	4.000	3.976	3.956
7	3.726	3.677	3.637	3.603	3.575	3.550	3.529
8	3.438	3.388	3.347	3.313	3.284	3.259	3.237
9	3.230	3.179	3.137	3.102	3.073	3.048	3.025
10	3.072	3.020	2.978	2.943	2.913	2.887	2.865
100	2.032	1.975	1.927	1.886	1.850	1.819	1.792
700	1.952	1.893	1.844	1.802	1.766	1.734	1.706
800	1.950	1.892	1.843	1.801	1.764	1.732	1.704
900	1.949	1.890	1.841	1.799	1.763	1.731	1.703
1000	1.948	1.889	1.840	1.798	1.762	1.730	1.702
∞	1.941	1.882	1.833	1.791	1.755	1.723	1.694

j) Dystrybuanta standardowego rozkładu normalnego.

	0	***0,01***	***0,02***	***0,03***	***0,04***	***0,05***	***0,06***	***0,07***	***0,08***	***0,09***
0	0,500	0,504	0,508	0,512	0,516	0,520	0,524	0,528	0,532	0,536
0,1	0,540	0,544	0,548	0,552	0,556	0,560	0,564	0,567	0,571	0,575
0,2	0,579	0,583	0,587	0,591	0,595	0,599	0,603	0,606	0,610	0,614
0,3	0,618	0,622	0,626	0,629	0,633	0,637	0,641	0,644	0,648	0,652
0,4	0,655	0,659	0,663	0,666	0,670	0,674	0,677	0,681	0,684	0,688
0,5	0,691	0,695	0,698	0,702	0,705	0,709	0,712	0,716	0,719	0,722
0,6	0,726	0,729	0,732	0,736	0,739	0,742	0,745	0,749	0,752	0,755
0,7	0,758	0,761	0,764	0,767	0,770	0,773	0,776	0,779	0,782	0,785
0,8	0,788	0,791	0,794	0,797	0,800	0,802	0,805	0,808	0,811	0,813
0,9	0,816	0,819	0,821	0,824	0,826	0,829	0,831	0,834	0,836	0,839
1	0,841	0,844	0,846	0,848	0,851	0,853	0,855	0,858	0,860	0,862
1,1	0,864	0,867	0,869	0,871	0,873	0,875	0,877	0,879	0,881	0,883
1,2	0,885	0,887	0,889	0,891	0,893	0,894	0,896	0,898	0,900	0,901
1,3	0,903	0,905	0,907	0,908	0,910	0,911	0,913	0,915	0,916	0,918
1,4	0,919	0,921	0,922	0,924	0,925	0,926	0,928	0,929	0,931	0,932
1,5	0,933	0,934	0,936	0,937	0,938	0,939	0,941	0,942	0,943	0,944
1,6	0,945	0,946	0,947	0,948	0,949	0,951	0,952	0,953	0,954	0,954
1,7	0,955	0,956	0,957	0,958	0,959	0,960	0,961	0,962	0,962	0,963
1,8	0,964	0,965	0,966	0,966	0,967	0,968	0,969	0,969	0,970	0,971
1,9	0,971	0,972	0,973	0,973	0,974	0,974	0,975	0,976	0,976	0,977
2	0,977	0,978	0,978	0,979	0,979	0,980	0,980	0,981	0,981	0,982
2,1	0,982	0,983	0,983	0,983	0,984	0,984	0,985	0,985	0,985	0,986
2,2	0,986	0,986	0,987	0,987	0,987	0,988	0,988	0,988	0,989	0,989
2,3	0,989	0,990	0,990	0,990	0,990	0,991	0,991	0,991	0,991	0,992
2,4	0,992	0,992	0,992	0,992	0,993	0,993	0,993	0,993	0,993	0,994
2,5	0,994	0,994	0,994	0,994	0,994	0,995	0,995	0,995	0,995	0,995
2,6	0,995	0,995	0,996	0,996	0,996	0,996	0,996	0,996	0,996	0,996
2,7	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997
2,8	0,997	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,998
2,9	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,999	0,999	0,999
3	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999

Zadanie 1.

Liczbę szkód K_i zgłaszanych przez kierowców w ciągu roku w pewnym portfelu ubezpieczeń AC (***dalej w zadaniach 2, 3 i 4 portfel ten będzie oznaczany przez PAC***) modelowano z wykorzystaniem uogólnionych modeli liniowych, uwzględniając następujące dwie zmienne objaśniające:

Zmienna	*Opis*
GrupaRyzyka	Grupa ryzyka. Zmienna jakościowa przyjmująca następujące kategorie: *A, B, C, D, E*. Kategorie określają grupy ryzyka utworzone z uwzględnieniem wieku i doświadczenia kierowcy.
Terytorium	Terytorium zamieszkania kierowcy. Zmienna jakościowa przyjmująca następujące kategorie: *T1, T2, T3, T4, T5, T6.*

Oszacowano dwa modele:

- M_Poisson. W modelu tym założono, że zmienna objaśniająca K_i ma rozkład Poissona. Przyjęto logarytmiczną funkcję wiążącą.
- M_NB. W modelu tym założono, że zmienna objaśniająca K_i ma rozkład ujemny dwumianowy. Przyjęto logarytmiczną funkcję wiążącą.

Uzyskano następujące wyniki:

	Model: M_Poisson			Model: M_NB		
Effect	*Estimate*	*Std. Error*	*p-Value*	*Estimate*	*Std. Error*	*p-Value*
(Intercept)	-2.67063	0.04839	< 2e-16	-2.67209	0.04913	< 2e-16
GrupaRyzyka B	0.30035	0.15795	0.057219	0.30054	0.15988	0.060148
GrupaRyzyka C	1.14825	0.07106	< 2e-16	1.14456	0.07299	< 2e-16
GrupaRyzyka D	0.52041	0.08222	2.46e-10	0.51934	0.08352	5.04e-10
GrupaRyzyka E	-0.01456	0.05781	0.801200	-0.01409	0.05835	0.809187
Terytorium T1	-0.76410	0.07122	< 2e-16	-0.76269	0.07204	< 2e-16
Terytorium T2	-0.60942	0.06788	< 2e-16	-0.60868	0.06873	< 2e-16
Terytorium T3	-0.50677	0.07712	4.98e-11	-0.50625	0.07808	8.96e-11
Terytorium T4	-0.35930	0.06296	1.15e-08	-0.35961	0.06386	1.79e-08
Terytorium T5	-0.20817	0.06216	0.000811	-0.20807	0.06309	0.000974
	Null deviance: 16373 on 59999 degrees of freedom Residual deviance: 15987 on 59990 degrees of freedom log-likelihood: -10559.82			Null deviance: 15435 on 59999 degrees of freedom Residual deviance: 15060 on 59990 degrees of freedom log-likelihood: -10554.3 Theta: 2.540 Std. Err.: 0.865		

a) Dla obydwu modeli wyznaczyć kryteria informacyjne *AIC* i *BIC*. W oparciu o uzyskane wyniki wybrać lepszy model.
b) Wykorzystując wybrany model wyznaczyć różnicę między oszacowaną (dopasowaną) średnią roczną liczbą szkód w grupie kierowców najgorszych (powodujących średnio w ciągu roku najwięcej wypadków) i najlepszych (powodujących średnio w ciągu roku najmniej wypadków).

Odpowiedzi:

Odp. a)

Dla modelu M_Poisson:

$$AIC = 21139.64$$
$$BIC = 21229.67$$

Dla modelu M_NB

$$AIC = 21130.60$$
$$BIC = 21229.62$$

Biorąc pod uwagę wartości obliczonych kryteriów informacyjnych należało wskazać model M_NB.
Uwaga! Jeżeli został wskazany model M_Poisson (z odpowiednim uzasadnieniem) punkt ten też był zaliczany.

Odp. b)
Model M_NB: 0.1853

Ewentualnie dla modelu M_Poisson: 0.1864

Rozwiązanie:

Ad. a)

$$AIC = -2\ln(L) + 2k$$
$$BIC = -2\ln(L) + \ln(n) \cdot k$$

gdzie:
L – wartość funkcji wiarygodności dla modelu,
k – liczba parametrów modelu,
n – liczba obserwacji.
Zatem:
- dla modelu M_Poisson

$$AIC = -2 \cdot (-10559.82) + 2 \cdot 10 = 21139.64$$
$$BIC = -2 \cdot (-10559.82) + \ln(60000) \cdot 10 = 21229.67$$

- dla modelu M_NB

$$AIC = -2 \cdot (-10559.82) + 2 \cdot 11 = 21130.60$$
$$BIC = -2 \cdot (-10559.82) + \ln(60000) \cdot 11 = 21229.62$$

Ad. b)

Najlepszy kierowca:

Model	M_Poisson	M_NB
GrupaRyzyka	E	E
Terytorium	T1	T1

M_Poisson:

$$\hat{k}_{E,T1} = \exp(-2.67063 - 0.01456 - 0.76410) = 0.0318$$

M_NB:

$$\hat{k}_{E,T1} = \exp(-2.67209 - 0.01409 - 0.76269) = 0.0318$$

Najgorszy kierowca:

Model	M_Poisson	M_NB
GrupaRyzyka	C	C
Terytorium	T6	T6

M_Poisson:

$$\hat{k}_{C,T6} = \exp(-2.67063 + 1.14825) = 0.2182$$

M_NB:

$$\hat{k}_{C,T6} = \exp(-2.67209 + 1.14456) = 0.2171$$

Zadanie 2.

W trakcie konstrukcji modelu dla wysokości pojedynczego roszczenia Y_i (*severity model*) dla portfela *PAC*, brano pod uwagę te same zmienne objaśniające, jak w przypadku modelowania rocznej liczby szkód, czyli:

Zmienna	*Opis*
GrupaRyzyka	Grupa ryzyka. Zmienna jakościowa przyjmująca następujące kategorie: *A, B, C, D, E*. Kategorie określają grupy ryzyka utworzone z uwzględnieniem wieku i doświadczenia kierowcy.
Terytorium	Terytorium zamieszkania kierowcy. Zmienna jakościowa przyjmująca następujące kategorie: *T1, T2, T3, T4, T5, T6.*

Oszacowano trzy grupy modeli:

- Grupa 1: Stanowią ją modele GAM0, GAM1 i GAM2, w których przyjęto **rozkład gamma** dla zmiennej objaśnianej Y_i oraz logarytmiczną funkcję wiążącą.
- Grupa 2: Stanowią ją modele IG0, IG1 i IG2, w których przyjęto rozkład **odwrotny Gaussa** (*inverse Gaussian distribution*) dla Y_i oraz funkcję wiążącą postaci: $f(x) = \frac{1}{x^2}$.
- Grupa 3: Stanowią ją modele LN0, LN1 i LN2, w których dla zmiennej objaśnianej $\ln(Y_i)$ przyjęto **rozkład Gaussa** i identycznościową funkcję wiążącą.

Uzyskano następujące wyniki:

Model	GAM0		GAM1		GAM2	
Effect	*Estimate*	*p-Value*	*Estimate*	*p-Value*	*Estimate*	*p-Value*
(Intercept)	7.98889	< 2e-16	7.9571	<2e-16	7.9632	<2e-16
GrupaRyzyka B			-0.3095	0.2110	-0.3228	0.1923
GrupaRyzyka C			0.2866	0.0140	0.2936	0.0118
GrupaRyzyka D			0.0412	0.7560	0.0444	0.7374
GrupaRyzyka E			0.0558	0.5460	0.0665	0.4716
Terytorium T1					0.0336	0.7687
Terytorium T2					-0.1368	0.2089
Terytorium T3					0.0671	0.5882
Terytorium T4					-0.0078	0.9385
Terytorium T5					0.0090	0.9277
Parametr dyspersji	2.484192		2.45546		2.449639	
	Null deviance: 3529.3 on 2531 degrees of freedom Residual deviance: 3529.3 on 2531 degrees of freedom AIC: 45518		Null deviance: 3529.3 on 2531 degrees of freedom Residual deviance: 3508.8 on 2527 degrees of freedom AIC: 45508		Null deviance: 3529.3 on 2531 degrees of freedom Residual deviance: 3500.3 on 2522 degrees of freedom AIC: 45510	

Model	IG0		IG1		IG2	
Effect	*Estimate*	*p-Value*	*Estimate*	*p-Value*	*Estimate*	*p-Value*
(Intercept)	1.162e-07	< 2e-16	1.226e-07	< 2e-16	1.205e-07	1.28e-11
GrupaRyzyka B			1.051e-07	0.2669	1.065e-07	0.2607
GrupaRyzyka C			-5.350e-08	0.0063	-5.224e-08	0.0075
GrupaRyzyka D			-9.699e-09	0.7486	-1.022e-08	0.7358
GrupaRyzyka E			-1.294e-08	0.5345	-1.295e-08	0.5341
Terytorium T1					-7.877e-09	0.7509
Terytorium T2					3.034e-08	0.2665
Terytorium T3					-1.344e-08	0.6066
Terytorium T4					4.285e-09	0.8524
Terytorium T5					-1.074e-09	0.9619
Parametr dyspersji:	0.000842661		0.0008369857		0.0008396223	
	Null deviance: 2.3122 on 2531 degrees of freedom Residual deviance: 2.3122 on 2531 degrees of freedom AIC: 44859		Null deviance: 2.3122 on 2531 degrees of freedom Residual deviance: 2.3054 on 2527 degrees of freedom AIC: 44859		Null deviance: 2.3122 on 2531 degrees of freedom Residual deviance: 2.3030 on 2522 degrees of freedom AIC: 44867	

Model	LN0		LN1		LN2	
Effect	*Estimate*	*p-Value*	*Estimate*	*p-Value*	*Estimate*	*p-Value*
(Intercept)	7.2920	< 2e-16	7.2747	< 2e-16	7.2707	< 2e-16
GrupaRyzyka B			-0.0586	0.7459	-0.0667	0.7128
GrupaRyzyka C			0.2665	0.0018	0.2636	0.0020
GrupaRyzyka D			0.0085	0.9299	0.0108	0.9112
GrupaRyzyka E			-0.0255	0.7054	-0.0248	0.7137
Terytorium T1					0.0386	0.6444
Terytorium T2					-0.0866	0.2773
Terytorium T3					0.0434	0.6323
Terytorium T4					0.0238	0.7476
Terytorium T5					0.0139	0.8485
Parametr dyspersji:	1.314687		1.311269		1.31216	
	Null deviance: 3327.5 on 2531 degrees of freedom Residual deviance: 3327.5 on 2531 degrees of freedom AIC: 7881.3		Null deviance: 3327.5 on 2531 degrees of freedom Residual deviance: 3313.6 on 2527 degrees of freedom AIC: 7878.7		Null deviance: 3327.5 on 2531 degrees of freedom Residual deviance: 3309.3 on 2522 degrees of freedom AIC: 7885.4	

Na podstawie podanych wyników:

a) Z każdej grupy modeli wybrać najlepszy. Wybór uzasadnić (w uzasadnieniu należy **między innymi** posłużyć się odpowiednim testem).
b) Wykorzystując wybrane modele wyznaczyć prognozy wysokości pojedynczego roszczenia dla kierowcy z grupy ryzyka *A* i terytorium *T1*.

Odpowiedzi:

Odp. a)

Z **grupy 1** należało wskazać model: GAM1
Z **grupy 2** należało wskazać model: IG0
Z **grupy 3** należało wskazać model: LN1

.....

Odp. b)

Grupa 1
Model GAM1: $\hat{y}_{A,T1} = 2855.78$

Grupa 2
Model IG0: $\hat{y}_{A,T1} = 2933.57$

Grupa 3
Model LN1: $\hat{y}_{A,T1} = 2780.36$

Rozwiązanie:

Ad. a)
Odpowiedź powinna uwzględniać następujące elementy:

- Wskazanie najlepszego modelu w każdej grupie.
- Uzasadnienie, w którym:
  - porównuje się AIC po uwzględnieniu kolejnych zmiennych objaśniających,
  - porównuje się istotność redukcji dewiancji po uwzględnieniu kolejnych zmiennych objaśniających z wykorzystaniem testu *F* (ze względu na oszacowania parametru dyspersji), którego statystyka wyraża się wzorem:

$$F = \frac{D\left(y; \hat{\theta}^P\right) - D\left(y; \hat{\theta}^Q\right)}{\hat{\phi}(q-p)},$$

gdzie:
$D\left(y; \hat{\theta}^P\right)$ – dewiancja modelu o mniejszej liczbie parametrów p,
$D\left(y; \hat{\theta}^Q\right)$ – dewiancja modelu o większej liczbie parametrów q,
$\hat{\phi}$ - oszacowanie parametru dyspersji modelu o większej liczbie parametrów.

Uwaga ! Akceptowane było także wykorzystanie testu opartego na statystyce χ^2.

Grupa 1

Poniżej podano sposób wykorzystania testu F w przypadku porównania modeli GAM0 i GAM1:

- Obliczamy wartość statystyki

$$F = \frac{D(y;\hat{\theta}^P) - D(y;\hat{\theta}^Q)}{\hat{\phi}(q-p)} = \frac{3529.3 - 3508.8}{2.45546 \cdot 4} = 2.087$$

- Wartość krytyczna: $F_{kr} = F_{4,2531} = 2.374$
- Nie ma podstaw do odrzucenia hipotezy o nieistotnej redukcji dewiancji.

Wartości statystyki F w porównaniu:

- GAM1 i GAM2: $F = 0.694$
- GAM0 i GAM2: $F = 1.316$

Wniosek: Biorąc pod uwagę test F i poziom istotności 0.05 należało wybrać model GAM0. Uwzględniając jednak wartość kryterium AIC oraz, to że wartość statystyki F w porównaniu GAM0 i GAM1 jest bliska wartości krytycznej (co może wskazywać istotną redukcję dewiancji na poziomie 0,10) należało wskazać model GAM1.
Uwaga! Można było również wskazać model GAM0.

Grupa 2

Wartości statystyki F w porównaniu:

- IG0 i IG1: $F = 2.031$
- IG1 i IG2: $F = 0.578$
- IG0 i IG2: $F = 1.218$

Wniosek: Biorąc pod uwagę test F i poziom istotności 0.05 należało wybrać model IG0. Wybór ten potwierdza także taka sama wartość AIC dla obydwu modeli.
Uwaga! Można było również wskazać model IG1.

Grupa 3

Wartości statystyki F w porównaniu:

- LN0 i LN1: $F = 2.650$
- LN1 i LN2: $F = 0.657$
- LN0 i LN2: $F = 1.542$

Wniosek: Biorąc pod uwagę test F i poziom istotności 0.05 należało wybrać model LN1. Wybór ten potwierdza wartość AIC, która jest najniższa dla tego modelu.

Ad. b)
Grupa 1
Model GAM1: $\hat{y}_{A,T1} = \exp(7.9571) = 2855.78$

Grupa 2
Model IG0: $\hat{y}_{A,T1} = \sqrt{\frac{1}{0.0000001162}} = 2933.57$

Grupa 3
Model LN1: $\hat{y}_{A,T1} = \exp\left(7.2747 + \frac{1.311269}{2}\right) = 2780.36$

Zadanie 3.

Dla portfela *PAC* przeprowadzono taryfikację jednomodelową z wykorzystaniem uogólnionych modeli liniowych. Przyjęto za zmienną objaśnianą Y_i iloraz całkowitych roszczeń dla *i*-tej polisy do ekspozycji i założono, że jej rozkład należy do rodziny rozkładów Tweediego (*Tweedie family of distributions*). Brano pod uwagę te same zmienne objaśniające, jak w przypadku modelowania rocznej liczby szkód i wysokości pojedynczej szkody, czyli:

Zmienna	*Opis*
GrupaRyzyka	Grupa ryzyka. Zmienna jakościowa przyjmująca następujące kategorie: *A, B, C, D, E*. Kategorie określają grupy ryzyka utworzone z uwzględnieniem wieku i doświadczenia kierowcy.
Terytorium	Terytorium zamieszkania kierowcy. Zmienna jakościowa przyjmująca następujące kategorie: *T1, T2, T3, T4, T5, T6.*

Oszacowano trzy modele: TWEED0, TWEED1 i TWEED2, w których przyjęto logarytmiczną funkcję wiążącą. Uzyskano następujące wyniki:

Model	TWEED0		TWEED1		TWEED2	
Effect	*Estimate*	*p-Value*	*Estimate*	*p-Value*	*Estimate*	*p-Value*
(Intercept)	5.0154	< 2e-16	4.9053	< 2e-16	5.2934	< 2e-16
GrupaRyzyka B			-0.0728	0.8324	-0.0348	0.9209
GrupaRyzyka C			1.3582	0.0000	1.4052	< 2e-16
GrupaRyzyka D			0.6010	0.0007	0.5564	0.0020
GrupaRyzyka E			0.0011	0.9919	0.0518	0.6344
Terytorium T1					-0.7238	0.0000
Terytorium T2					-0.7300	0.0000
Terytorium T3					-0.4448	0.0040
Terytorium T4					-0.3631	0.0060
Terytorium T5					-0.1622	0.2196
Parametr dyspersji (*dispersion parameter*)	429.450		430.800		430.890	
Indeks (index parameter)	1.5484		1.5422		1.5389	
AIC	66814		66692		66619	

a) W oparciu o przedstawione w powyższej tabeli wyniki wybrać najlepszy model (**wybór uzasadnić**).
b) Zinterpretować parametry wybranego modelu.
c) Wyznaczyć różnicę między najwyższą i najniższą wartością $\hat{Y}_i$ otrzymaną z wykorzystaniem wybranego w punkcie a) modelu.

Odpowiedzi:

.....

Odp. a)

Należało wskazać model TWEED2.

.....

Odp. b)

Interpretujemy tylko parametry statystycznie istotne, przyjmując poziom istotności równy 0.05.

- Wyraz wolny wynosi 5.2934 i jest statystycznie istotny. Jest to wartość predyktora liniowego dla bazowej grupy kierowców, tzn. z grupy ryzyka A i terytorium T6. Wartość $\exp(5.2934) = 199.02$ jest prognozowaną wartością składki czystej (*pure premium*) dla tej grupy kierowców.
- Parametr stojący przy kategorii *GrupaRyzyka:*C wynosi 1.4052 i jest statystycznie istotny. Wskazuje on, że składka czysta dla kierowców z grupy ryzyka C jest o $(\exp(1.4052) - 1) * 100\% = 307.63\%$ wyższa w porównaniu z kierowcami z grupy ryzyka A, przy założeniu zasady ceteris paribus (pozostałe zmienne pozostają bez zmian).
- Parametr stojący przy kategorii *GrupaRyzyka:*D wynosi 0.5564 i jest statystycznie istotny. Wskazuje on, że składka czysta dla kierowców z grupy ryzyka C jest o $(\exp(0.5564) - 1) * 100\% = 74.44\%$ wyższa w porównaniu z kierowcami z grupy ryzyka A, przy założeniu zasady ceteris paribus (pozostałe zmienne pozostają bez zmian).
- Parametr stojący przy kategorii *Terytorium:*T1 wynosi -0.7238 i jest statystycznie istotny. Wskazuje on, że składka czysta dla kierowców z terytorium T1 jest o $(1 - \exp(-0.7238)) * 100\% = 51.51\%$ niższa w porównaniu z kierowcami z terytorium T6, przy założeniu zasady ceteris paribus (pozostałe zmienne pozostają bez zmian).

Analogicznie interpretujemy pozostałe parametry.

.....

Odp. c)

Różnica: $811.27 - 92.63 = 718.64$

Rozwiązanie:

Ad. a)

Odpowiedź powinna uwzględniać następujące elementy:

- Wskazanie najlepszego modelu.
- Uzasadnienie, w którym porównuje się AIC po uwzględnieniu kolejnych zmiennych objaśniających,

Ad. b) Patrz odpowiedź do b).

Ad. c)
Dla modelu TWEED2:

Najwyższa wartość:

Model	TWEED2
GrupaRyzyka	C
Terytorium	T6

$\hat{y}_{C,T6} = \exp(5.2934 + 1.4052) = 811.27$

Najniższa wartość:

Model	TWEED2
GrupaRyzyka	B
Terytorium	T2

$\hat{y}_{B,T2} = \exp(5.2934 - 0.0348 - 0.7300) = 92.63$

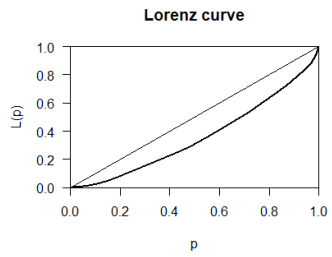
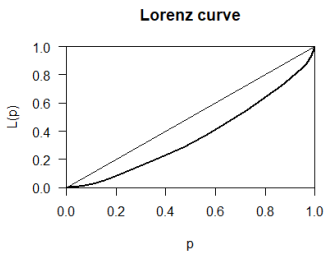
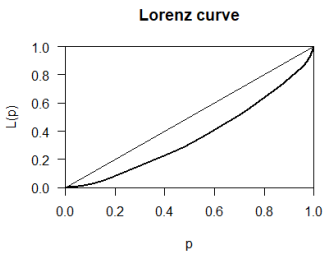
Zadanie 4.

Dla portfela *PAC* rozważano trzy plany taryfikacji:

- **Plan_A**: Taryfikację dwumodelową, w której liczba szkód jest modelowana za pomocą podanego w **zadaniu 1** modelu M_NB, natomiast wysokość pojedynczego roszczenia - podanego w **zadaniu 2** modelu GAM2.
- **Plan_B**: Taryfikację dwumodelową, w której liczba szkód jest modelowana za pomocą podanego w **zadaniu 1** modelu M_NB, natomiast wysokość pojedynczego roszczenia - podanego w **zadaniu 2** modelu LN2.
- **Plan_C:** Taryfikację jednomodelową z wykorzystaniem podanego w **zadaniu 3** modelu TWEED2.

Plany te przetestowano na zbiorze testowym liczącym 40000 obserwacji. Uzyskano następujące wyniki:

- Krzywe Lorenza:

Plan_A	**Plan_B**	**Plan_C**
		
Pole pod krzywą Lorenza		
0.3572552	0.3595529	0.3580689

- Wskaźnik $w = \frac{\sum_{i=1}^{n} \hat{z}_i}{\sum_{i=1}^{n} z_i}$, gdzie $\hat{z}_i$ i z_i oznaczają składki czyste (*pure premiums*) odpowiednio oszacowane i ze zbioru testowego.

Plan	Wskaźnik *w* dla całego zbioru testowego
Plan_A	0.9288364
Plan_B	0.8943979
Plan_C	0.9377649

Wskaźnik *w* dla obserwacji w zbiorze testowym z grupy ryzyka:	Plan		
	Plan_A	**Plan_B**	**Plan_C**
A	0.8752	0.8550	0.8876
B	0.4618	0.5834	0.4619
C	1.0949	1.0381	1.0732
D	0.9349	0.8826	0.9407
E	1.2349	1.1020	1.2522

a) Wyznaczyć składkę czystą dla kierowcy z grupy ryzyka *A* i terytorium *T1*, wykorzystując każdy z przedstawionych planów.
b) Omówić możliwość wykorzystania krzywej Lorenza i współczynnika Giniego w taryfikacji.
c) Wyznaczyć współczynnik Giniego dla każdego planu taryfikacji.
d) Na podstawie podanych w tym zadaniu informacji wybrać najlepszy plan taryfikacji. **Wybór uzasadnić.**

Odpowiedzi:

.....

Odp. a)

Plan A: $P = 95.680$
Plan B: $P = 92.721$
Plan C: $P = 96.505$

.....

Odp. b)

Odpowiedź powinna uwzględniać następujące elementy:

- Omówienie ogólnego zastosowania krzywej Lorenza i współczynnika Giniego.
- Ogólne przedstawienie sposobu konstrukcji krzywej Lorenza na potrzeby oceny planów taryfikacji.
- Wskazanie, że krzywa Lorenza i współczynnik Giniego pomaga w wyborze planów taryfikacji lepiej różnicujących ryzyko w portfelu.

.....

Odp. c)

Plan A: $G_A = 0.2854896$
Plan B: $G_B = 0.2808942$
Plan C: $G_C = 0.28386222$

.....

Odp. d)

Odpowiedź powinna uwzględniać następujące elementy:

- Wskazanie planu C (ewentualnie planu A).
- Uzasadnienie wyboru, odwołujące się do współczynnika Giniego i wskaźników rentowności *w*.

Rozwiązanie:

Ad. a)
Dane:

GrupaRyzyka	A
Terytorium	T1

$\hat{k}_{A,T1} = \exp(-2.67209 - 0.76269) = 0.0322$

Składka czysta:
Plan A:

$$P = \hat{k}_{A,T1} \cdot \hat{y}_{A,T1}$$
$$\hat{y}_{A,T1} = \exp(7.9632 + 0.0336) = 2971.434$$
$$P = 0.0322 \cdot 2971.434 = 95.680$$

Plan B:

$$P = \hat{k}_{A,T1} \cdot \hat{y}_{A,T1}$$

$$\hat{y}_{A,T1} = \exp\left(7.2747 + 0.0386 + \frac{1.31216}{2}\right) = 2879.523$$
$$P = 0.0322 \cdot 2879.523 = 92.721$$

Plan C:

$$P = \exp(5.2934 - 0.7238) = 96.505$$

Ad. b) Patrz odpowiedź do b).

Ad. c)
Współczynnik Giniego

$$G = 2 \cdot (0.5 - pole\ pod\ krzywą\ Lorenza)$$

Plan A:

$$G_A = 2 \cdot (0.5 - 0.3572552) = 0.2854896$$

Plan B:

$$G_B = 2 \cdot (0.5 - 0.3595529) = 0.2808942$$

Plan C:

$$G_C = 2 \cdot (0.5 - 0.3580689) = 0.28386222$$

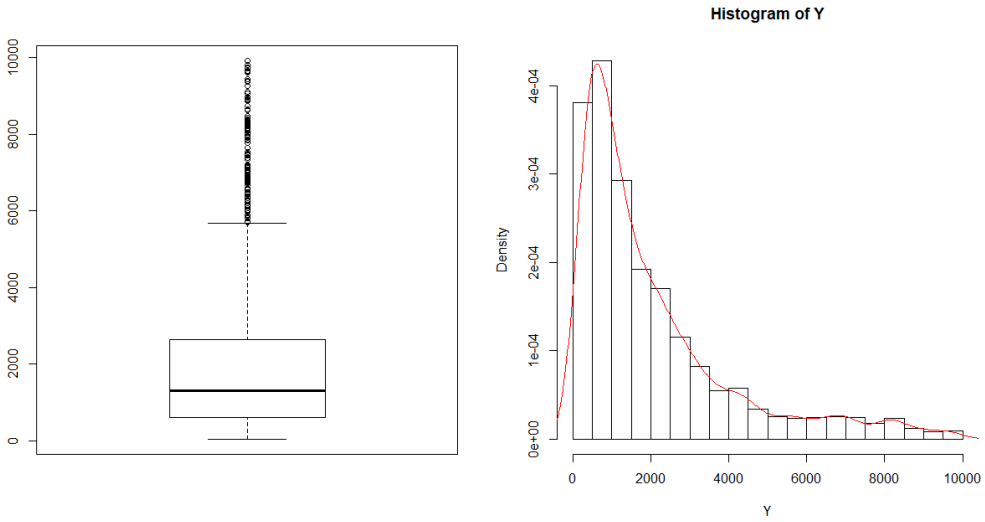
Ad. d) Patrz odpowiedź do d).

Zadanie 5.

Identyfikację rozkładu zmiennej losowej Y modelującej wysokość pojedynczego roszczenia w pewnym portfelu ubezpieczeń aktuariusz przeprowadził w dwóch etapach.

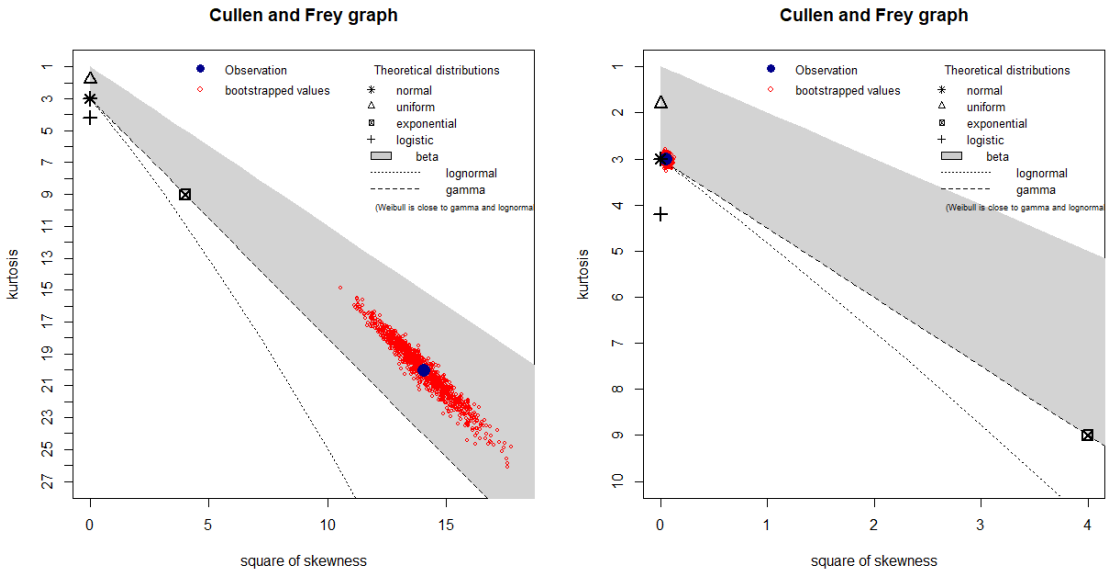
1. Etap 1. Sporządził zamieszczone poniżej rysunki: Rys.1/zad.5 i Rys.2/zad.5. (Z lewej strony na Rys.2/zad.5 przedstawił wykres Cullena-Freya dla Y, z prawej dla $\ln(Y)$.) Na tej podstawie wstępnie wybrał dwa rozkłady, tj. rozkład logarytmiczno-normalny i gamma.
2. Etap 2. Oszacował parametry wybranych w etapie pierwszym rozkładów, uzyskując dystrybuanty teoretyczne: $F_{LogNorm}$ i F_{Gamma}. Następnie sporządził przedstawione na Rys.3/zad.5 histogramy zmiennych $F_{LogNorm}(Y)$ i $F_{Gamma}(Y)$ (na wykresach oznaczone odpowiednio F.Log-Norm i F.Gamma) i na tej podstawie ostatecznie wybrał rozkład logarytmiczno-normalny.

Rys.1/zad.5

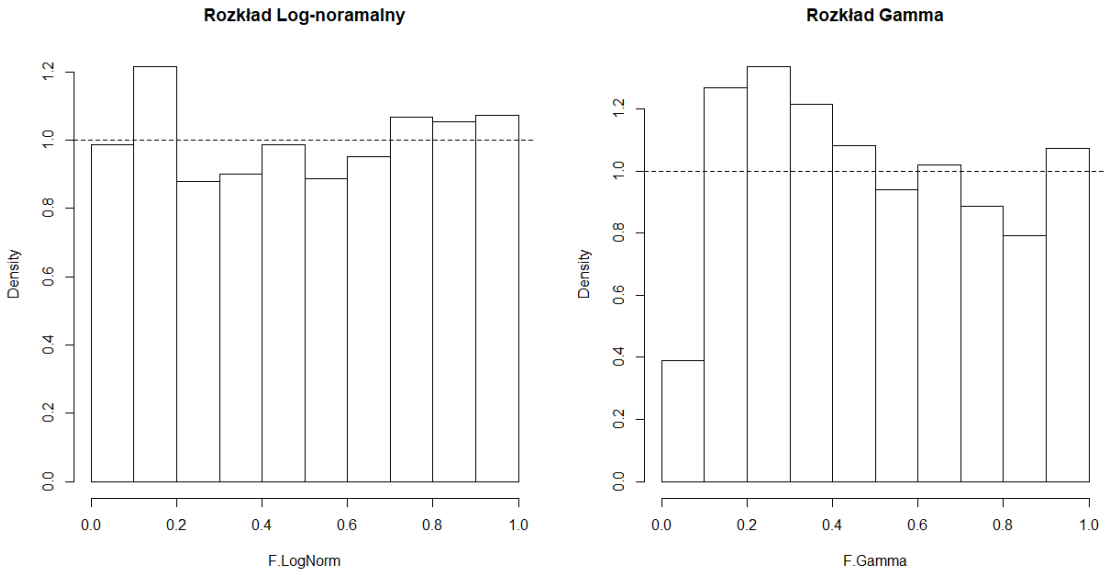


Rys.2/zad.5

Y $\ln(Y)$



Rys.3/zad.5



a) Przedstawić sposób konstrukcji i wykorzystania wykresu Cullena-Freya.
b) Czy postępowanie aktuariusza na pierwszym i drugim etapie wyboru modelu było poprawne. **Odpowiedź uzasadnić.**

Odpowiedzi:

.....

Odp. a)

Odpowiedź powinna uwzględniać następujące elementy:

- Opis osi układu współrzędnych.
- Opis elementów składających się na wykres Cullena-Freya.
- Informację, że wykres Cullena-Freya można wykorzystać do wstępnego wyboru rodziny rozkładów ciągłych w modelowaniu wysokości roszczeń.

.....

Odp. b)

Odpowiedź powinna uwzględniać następujące elementy:

- Informację, że postępowanie aktuariusza na pierwszym i drugim etapie wyboru modelu było poprawne.
- Uzasadnienie poprawności postępowania aktuariusza na pierwszym etapie, odwołujące się do zamieszczonych wykresów na Rys. 1/5 i Rys.2/5 i do faktu, że jeżeli zmienna losowa X ma rozkład logarytmiczno-normalny, to $\ln(X)$ ma rozkład normalny.
- Uzasadnienie poprawności postępowania aktuariusza na drugim etapie, odwołujące się do zamieszczonych wykresów na Rys. 3/5 i faktu, że jeżeli ciągła zmienna losowa X posiada dystrybuantę F, to $F(X) \sim U_{(0,1)}$, gdzie $U_{(0,1)}$ oznacza rozkład jednostajny na przedziale $(0,1)$.

Rozwiązanie:

Ad. a) Patrz odpowiedź a).

Ad. b) Patrz odpowiedź b).

Zadanie 6.

Rozważano portfel składający się z polis ubezpieczeniowych zapewniających ochronę przed dwoma rodzajami ryzyka: R_1 i R_2. Przyjęto, że dla i-tej polisy wektor losowy $Y_i = (Y_{i1}, Y_{i2})$ modeluje koszty roszczeń odpowiednio dla pierwszego i drugiego ryzyka. Dla i –tej polisy z tego portfela wystąpienie roszczenia związanego z realizacją ryzyka R_1 i R_2, modelowano wykorzystując następujący wektor losowy:

$$K_i = (K_{i1}, K_{i2}),$$

gdzie $K_{i1} = I(Y_{i1} > 0)$, $K_{i2} = I(Y_{i2} > 0)$, przy czym:

$$I(X > 0) = \begin{cases} 1, gdy\ X > 0 \\ 0, gdy\ X \leq 0 \end{cases}.$$

Założono, że rozkład $K_i = (K_{i1}, K_{i2})$ jest taki sam dla każdej polisy (w dalszym ciągu będzie oznaczany bez indeksu i, czyli $K = (K_1, K_2)$) i oszacowano jego rozkłady brzegowe, otrzymując następujące wyniki:

$$P(K_1 = 0) = 0.95, \qquad P(K_1 = 1) = 0.05$$
$$P(K_2 = 0) = 0.80, \qquad P(K_2 = 1) = 0.20$$

Wyznaczyć rozkład wektora K, zakładając, że:

a) Zmienne K_1, K_2 są niezależne.

b) Struktura zależności między zmiennymi K_1, K_2 jest modelowana za pomocą kopuli Claytona:

$$C(u, v) = \left(u^{-\theta} + v^{-\theta} - 1\right)^{-\frac{1}{\theta}},$$

dla której $\hat{\theta} = 1$.

Odpowiedzi:

.....

Odp. a)

$$P(K_1 = 1, K_2 = 1) = 0.01$$
$$P(K_1 = 1, K_2 = 0) = 0.04$$
$$P(K_1 = 0, K_2 = 1) = 0.19$$
$$P(K_1 = 0, K_2 = 0) = 0.76$$

.....

Odp. b)

$$P(K_1 = 1, K_2 = 1) = 0.0177$$
$$P(K_1 = 1, K_2 = 0) = 0.0323$$
$$P(K_1 = 0, K_2 = 1) = 0.1823$$
$$P(K_1 = 0, K_2 = 0) = 0.7677$$

Rozwiązanie:

Ad. a)
W przypadku, gdy zmienne K_1, K_2 są niezależne:

$$P(K_1 = 1, K_2 = 1) = 0.05 \cdot 0.20 = 0.01$$
$$P(K_1 = 1, K_2 = 0) = 0.05 \cdot 0.80 = 0.04$$
$$P(K_1 = 0, K_2 = 1) = 0.95 \cdot 0.20 = 0.19$$
$$P(K_1 = 0, K_2 = 0) = 0.95 \cdot 0.80 = 0.76$$

Ad. b)
W przypadku, gdy struktura zależności między zmiennymi K_1, K_2 jest modelowana za pomocą kopuli C:
$P(K_1 = 1, K_2 = 1) = 1 - F_1(0) - F_2(0) + C\big(F_1(0), F_2(0)\big)$,
$P(K_1 = 1, K_2 = 0) = F_2(0) - C\big(F_1(0), F_2(0)\big)$,
$P(K_1 = 0, K_2 = 1) = F_1(0) - C\big(F_1(0), F_2(0)\big)$,
$P(K_1 = 0, K_2 = 0) = C\big(F_1(0), F_2(0)\big)$,
gdzie F_1 i F_2 to dystrybuanty odpowiednio zmiennych K_1, K_2 .

Stąd:
$F_1(0) = 0.95$,
$F_2(0) = 0.80$,
$C\big(F_1(0), F_2(0)\big) = ((0.95)^{-1} + (0.80)^{-1})^{-1} = 0.7677$.

I ostatecznie:
$P(K_1 = 1, K_2 = 1) = 0.0177$
$P(K_1 = 1, K_2 = 0) = 0.0323$,
$P(K_1 = 0, K_2 = 1) = 0.1823$,
$P(K_1 = 0, K_2 = 0) = 0.7677$.

Zadanie 7.

a) Na czym polegają i czemu służą procedury agregujące rodziny klasyfikatorów (*ensemble methods*)?
b) Przedstawić co najmniej dwie takie procedury.

Odpowiedzi:

.....

Odp. a)

Odpowiedź powinna uwzględniać następujące elementy:

- Opis głównej idei metody klasyfikatorów zespołowych.
- Informację w jakim celu są wykorzystywane.

.....

Odp. b)

Odpowiedź powinna uwzględniać następujące elementy:

- Wskazanie dwóch procedur agregujących (np. *baggin*, *boosting*).
- Krótką charakterystykę tych procedur.

Rozwiązanie:

Ad. a) Patrz odpowiedź a).

Ad. b) Patrz odpowiedź b).

Zadanie 8.

Odnowienie przez kierowców polisy AC w pewnej firmie ubezpieczeniowej XYZ modelowano z wykorzystaniem regresji probitowej (uogólnionego modelu liniowego z funkcją wiążącą *probit*). Przyjęto, że zmienna zależna Y_i przyjmuje dwie wartości:
$Y_i = 1$, gdy kierowca przedłuży umowę na kolejny rok
$Y_i = 0$, gdy nie przedłuży.
Oszacowano dwa modele:
Model A. W modelu tym uwzględniono następujące zmienne objaśniające:

- *plec:* Płeć (K - kobieta. M – mężczyzna)
- *zamieszkanie:* Miejsce zamieszkania (Miasto – miasto. Wies – wieś)
- *stan.cywilny:* Stan cywilny (C – małżeństwo. S – singiel. O – inny)
- *wiek:* Wiek kierowcy w latach.
- *lojalnosc:* Liczba lat, w których kierowca był klientem firmy XYZ.

Na podstawie zbioru uczącego liczącego 1865 obserwacji, metodą największej wiarygodności uzyskano następujące oszacowania:

Coefficients:

	Estimate	*Std. Error*	*p-Value*
(Intercept)	1.1485	0.1379	<2e-16
*plec*_M	-0.7383	0.0745	<2e-16
*zamieszkanie*_Wies	-0.7350	0.0682	<2e-16
*stan.cywilny*_C	-0.0091	0.0873	0.9170
*stan.cywilny*_O	0.2284	0.1767	0.1960
wiek	-0.0326	0.0033	<2e-16
lojalnosc	0.0806	0.0059	<2e-16

Null deviance: 2380.2 on 1864 degrees of freedom
Residual deviance: 1937.5 on 1858 degrees of freedom
AIC: 1951.5

Dla tego modelu tabela analizy dewiancji (*Deviance Table*) przedstawia się następująco:

Model: binomial. link: probit
Response: Y
Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			1864	2380.2	
plec	1	101.075	1863	2279.1	< 2.2e-16
zamieszkanie	1	97.467	1862	2181.6	< 2.2e-16
stan.cywilny	2	9.176	1860	2172.4	0.01017
wiek	1	19.830	1859	2152.6	8.46E-06
lojalnosc	1	215.085	1858	1937.5	< 2.2e-16

Model B. W modelu tym wzięto pod uwagę te same zmienne co w modelu A, z tym, że zmienne ilościowe *wiek* i *lojalnosc* zamieniono na jakościowe. Uwzględniono zatem:

- *plec:* Płeć (K - kobieta. M – mężczyzna)
- *zamieszkanie:* Miejsce zamieszkania (Miasto – miasto. Wies – wieś)
- *stan.cywilny:* Stan cywilny (C – małżeństwo. S – singiel. O – inny)
- *wiek.kat:* Zmienna jakościowa przyjmująca trzy kategorie: 25 lat i mniej, (25, 60], powyżej 60 lat
- *lojalnosc.kat:* Zmienna jakościowa przyjmująca trzy kategorie: 2 lata i mniej, (2, 10], powyżej 10 lat.

Na podstawie zbioru uczącego liczącego 1865 obserwacji, metodą największej wiarygodności uzyskano następujące oszacowania:

Coefficients:

	Estimate	*Std. Error*	*p-Value*
(Intercept)	0.6269	0.2125	0.0032
*plec*_M	-0.7144	0.0733	< 2e-16
*zamieszkanie*_Wies	-0.6992	0.0665	< 2e-16
*stan.cywilny*_C	-0.1909	0.0836	0.0225
*stan.cywilny*_O	-0.0112	0.1724	0.9481
wiek.kat (25,60]	-0.5050	0.2012	0.0121
wiek.kat powyżej 60	-1.1172	0.2244	0.0000
lojalnosc.kat (2,10]	0.3064	0.1387	0.0272
l*ojalnosc.kat* powyżej10	0.9712	0.1410	0.0000

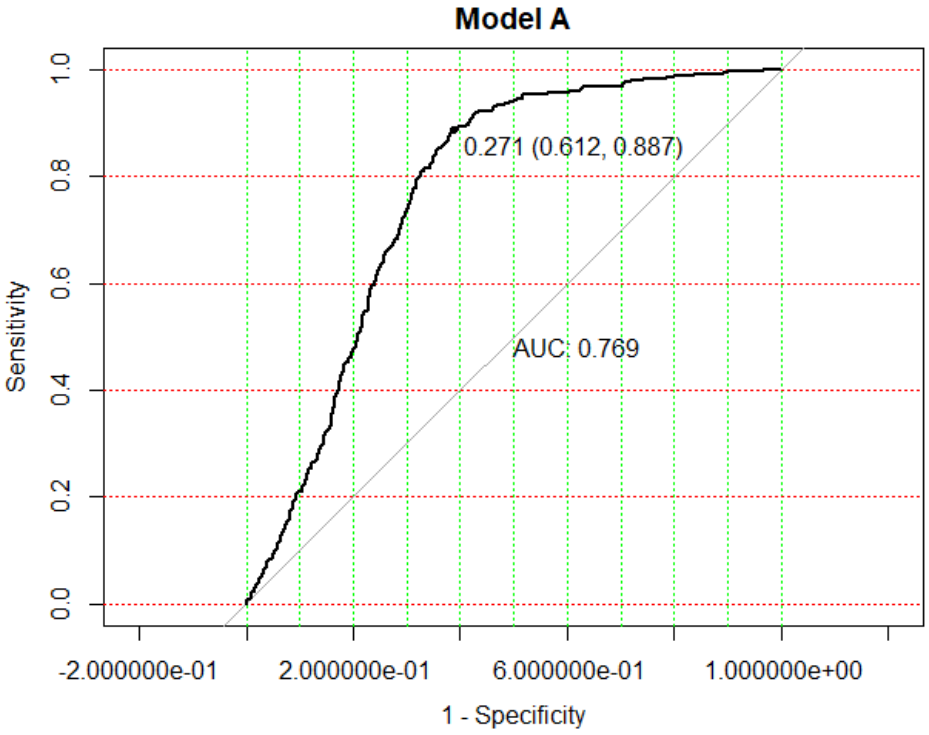
Null deviance: 2380.2 on 1864 degrees of freedom
Residual deviance: 2031.9 on 1856 degrees of freedom
AIC: 2049.9

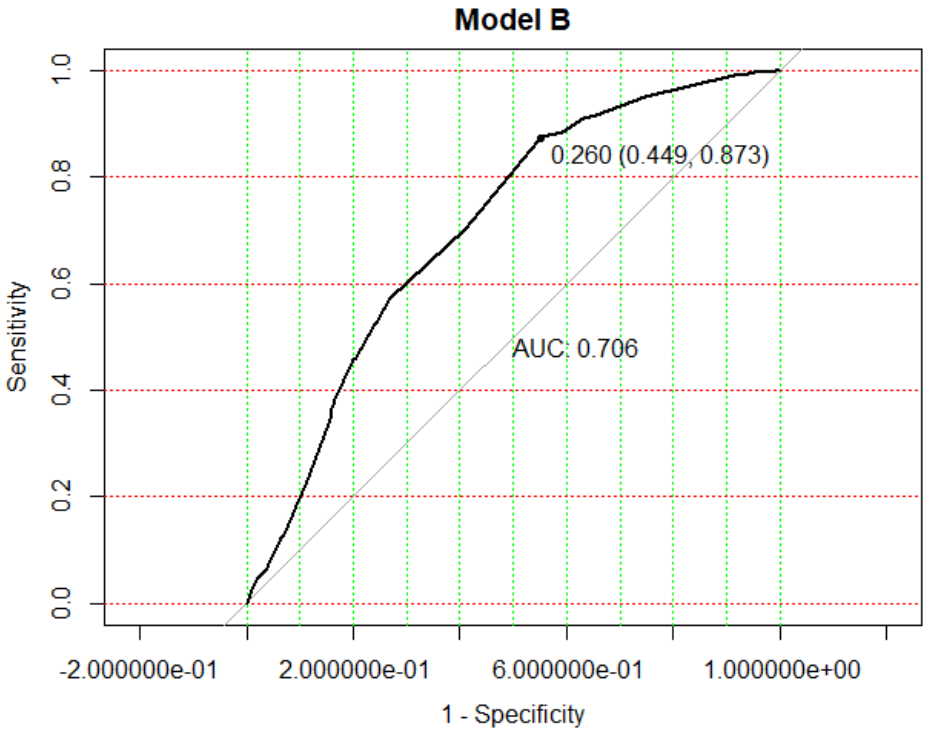
Dla tego modelu tabela analizy dewiancji (*Deviance Table*) przedstawia się następująco:

Model: binomial. link: probit
Response: Y
Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			1864	2380.2	
plec	1	101.075	1863	2279.1	< 2.2e-16
zamieszkanie	1	97.467	1862	2181.6	< 2.2e-16
stan.cywilny	2	9.176	1860	2172.4	0.0102
wiek.kat	2	25.461	1858	2147.0	0.0000
lojalnosc.kat	2	115.082	1856	2031.9	< 2.2e-16

Na podstawie zbioru testowego liczącego 1244 obserwacje wyznaczono dla tych modeli:
- krzywe ROC





- tabele trafności prognoz (punkty odcięcia dla poszczególnych modeli ustalono na podstawie odpowiednich krzywych ROC):

Model A:

		Stan faktyczny	
		$Y_i = 0$ (N)	$Y_i = 1$ (P)
Prognoza	$Y_i^P = 0$ (N)	505	47
	$Y_i^P = 1$ (P)	322	370

Model B:

		Stan faktyczny	
		$Y_i = 0$ (N)	$Y_i = 1$ (P)
Prognoza	$Y_i^P = 0$ (N)	371	53
	$Y_i^P = 1$ (P)	456	364

a) Za pomocą oszacowanych modeli A i B wyznaczyć prognozę zmiennej Y_i dla 65-cio letniego żonatego mężczyzny, mieszkającego na wsi, który był lojalnym klientem firmy XYZ przez 26 lat. Punkt odcięcia dla poszczególnych modeli ustalić na podstawie podanych wykresów krzywych ROC.
b) Dla rozważanych modeli A i B, na podstawie podanych wyników, podać specyficzność (*specificity*), czułość (s*ensitivity*), precyzję wskazania negatywnego oraz precyzję wskazania pozytywnego.
c) W oparciu podane informacje wybrać lepszy model. Wybór uzasadnić, biorąc pod uwagę informacje uzyskane zarówno na etapie estymacji, jak i testowania modeli.

Odpowiedzi:

.....

Odp. a)

Prognoza:

- Model A: $y_A^P = 1$
- Model A: $y_B^P = 0$

.....

Odp. b)

Model A:

$specyficzność = 0.612$ (można było odczytać z wykresu krzywej ROC)
$czułość = 0.887$ (można było odczytać z wykresu krzywej ROC)
$precyzja\ wskazania\ negatywnego = 0.915,$
$precyzja\ wskazania\ pozytywnego = 0.535.$

Model B:

$specyficzność = 0.449$ (można było odczytać z wykresu krzywej ROC)
$czułość = 0.873$ (można było odczytać z wykresu krzywej ROC)
$precyzja\ wskazania\ negatywnego = 0.875,$
$precyzja\ wskazania\ pozytywnego = 0.444.$

.....

Odp. c)

Odpowiedź powinna uwzględniać następujące elementy:

- Wskazanie modelu A.
- Uzasadnienie wyboru, odwołujące się zarówno do informacji uzyskanych na etapie estymacji, jak i testowania modeli.

Rozwiązanie:

Ad a)

Dane:

plec: M
zamieszkanie: Wies
stan.cywilny: C
wiek: 65
lojalnosc: 26

Model A:

$$\hat{\eta}_i = 1.1485 - 0.7383 - 0.7350 - 0.0091 - 0.0326 \cdot 65 + 0.0806 \cdot 26 = -0.3573$$

$$\hat{\eta}_i \approx -0.36$$

$$\hat{\mathrm{p}} = \Phi(-0.36) = 1 - \Phi(0.36) = 1 - 0.641 = 0.359$$

Punkt odcięcia: 0.271, otrzymujemy zatem $\hat{\mathrm{p}} > 0.271$, czyli $y_A^P = 1$.

Model B:

$$\hat{\eta}_i = 0.6269 - 0.7144 - 0.6992 - 0.1909 - 1.1172 + 0.9712 = -1.1236$$

$$\hat{\eta}_i \approx -1.12$$

$$\hat{\mathrm{p}} = \Phi(-1.12) = 1 - \Phi(1.12) = 1 - 0.869 = 0.131$$

Punkt odcięcia: 0.260, otrzymujemy zatem $\hat{\mathrm{p}} < 0.260$, czyli $y_B^P = 0$.

Ad. b)

Model A:

$specyficzność = \frac{505}{505+322} = 0.612$ (można było odczytać z wykresu krzywej ROC)

$czułość = \frac{370}{370+47} = 0.887$ (można było odczytać z wykresu krzywej ROC)

$precyzja\ wskazania\ negatywnego = \frac{505}{505+47} = 0.915,$

$precyzja\ wskazania\ pozytywnego = \frac{370}{370+322} = 0.535.$

Model B:

$specyficzność = \frac{371}{371+456} = 0.449$ (można było odczytać z wykresu krzywej ROC)

$czułość = \frac{364}{364+53} = 0.873$ (można było odczytać z wykresu krzywej ROC)

$precyzja\ wskazania\ negatywnego = \frac{371}{371+53} = 0.875,$

$precyzja\ wskazania\ pozytywnego = \frac{364}{364+456} = 0.444.$

Ad. c) Patrz odpowiedź do c)

Zadanie 9.

W celu przeprowadzenia taryfikacji oszacowano dwa zagnieżdżone uogólnione modele liniowe A i B. W modelu B w porównaniu z A uwzględniono dodatkowo jedną, nominalną zmienną objaśniającą X przyjmującą wartości ze zbioru liczącego 3 kategorie. Zbiór uczący liczył 2000 obserwacji. Uzyskano między innymi następujące informacje:

	Model nasycony (*Saturated Model*	Model A	Model B
Logarytm wiarygodności (*Log-Likelihood*)	-500	-1000	-980
Oszacowanie parametru dyspersji	3.0	3.0	3.0

a) Wykorzystując odpowiedni test i wybrane kryterium informacyjne sprawdzić, czy zmienną X należy uwzględnić w taryfikacji.
b) Wyjaśnić dlaczego w porównywaniu zagnieżdżonych uogólnionych modeli liniowych nie powinno się uwzględniać tylko dewiancji.

Odpowiedzi:

.....

Odp. a)
Zmienną X powinno się uwzględnić w taryfikacji

.....

Odp. b)
Odpowiedź powinna uwzględniać następujące elementy:

- Wskazanie, że dodanie kolejnej zmiennej do modelu wpływa na obniżenie dewiancji.
- Wskazanie, że wykorzystanie tylko dewiancji może prowadzić do budowy modelu z nadmierną liczbą parametrów oraz do jego przeuczenia.

Rozwiązanie:

Ad. a)
Korzystamy ze statystyki F (patrz rozwiązanie zad. 2):

$$F = \frac{D\left(y; \hat{\theta}^A\right) - D\left(y; \hat{\theta}^B\right)}{\hat{\phi}(q-p)},$$

gdzie q i p oznaczają liczbę parametrów odpowiednio modelu B i A.

Obliczenia:

$$D\left(y;\hat{\theta}^{A}\right) - D\left(y;\hat{\theta}^{B}\right) = 2(\ln(L_B) - \ln(L_A)) = 2\left(-980 - (-1000)\right) = 40$$

ponieważ:

$D\left(y;\hat{\theta}^{A}\right) = 2(\ln(L_{Sat}) - \ln(L_A))$,

$D\left(y;\hat{\theta}^{B}\right) = 2(\ln(L_{Sat}) - \ln(L_B))$,

gdzie L_{Sat} wiarygodność modelu nasyconego.

Stąd:

$$F = \frac{40}{3 \cdot 2} = 6.667.$$

Wartość krytyczna:

$$F_{kr} = F_{2,2000-p} \approx 2.998$$

Kryteria informacyjne:

$$AIC_A = -2\ln(L_A) + 2p = 2 \cdot 1000 + 2p = 2000 + 2p$$

$$AIC_B = -2\ln(L_B) + 2(p+2) = 2 \cdot 980 + 2p + 4 = 1964 + 2p$$

Czyli:

$$AIC_B < AIC_A$$

Ad. b) Patrz odpowiedź do b).

Zadanie 10.
Wiadomo, że:

- zmienna losowa X_1 modelująca straty dla pierwszej linii biznesu ma rozkład gamma z parametrami $\alpha = 2$ i $\beta = 50$;
- zmienna losowa X_2 modelująca straty dla drugiej linii biznesu ma rozkład wykładniczy z parametrem $\lambda = 0.02$;
- zmienna losowa X_3 modelująca straty dla trzeciej linii biznesu ma rozkład normalny z parametrami $\mu = 100$ i $\sigma = 20$.

Łączne wymogi kapitałowe dla tych trzech linii (czyli dla $S = X_1 + X_2 + X_3$) wyznaczono wykorzystując następujące metody:

- **Metoda M1.** Dla pojedynczych linii biznesu wyznaczono wymogi kapitałowe równe wartości zagrożonej na poziomie bezpieczeństwa $\alpha = 0.995$, a następnie zagregowano je wykorzystując standardową metodę wariancji-kowariancji (stosowaną w Solvency II). Przyjęto, że współczynnik korelacji liniowej między:
  - X_1 i X_2 jest równy 0.60,
  - X_1 i X_3 jest równy 0.50,
  - X_2 i X_3 jest równy 0.80.
- **Metoda M2.** Najpierw zidentyfikowano strukturę zależności między stratami X_1, X_2, X_3. Okazało się, że najlepiej opisuje ją kopula Gumbela z parametrem $\theta = 2$. Następnie wykorzystując metodę symulacji wyznaczono wymóg kapitałowy na podstawie następującej formuły: $\kappa = VaR_{0.995}(S) - E(S)$. Wartość zagrożoną $VaR_{0.995}(S)$ uzyskano na podstawie podanych w poniższej tabeli statystyk pozycyjnych $S_{(1081:1000)}. \ldots . S_{(1100:1100)}$ dla 1100 wartości zmiennej S uzyskanych w drodze symulacji. Wartości ***A***, ***B*** i ***C*** odpowiadają następującym ciągom wygenerowanym z zastosowanej kopuli:
  $(u_1^A, u_2^A, u_3^A) = (0.994,\ 0{,}986,\ 0{,}995)$,
  $(u_1^B, u_2^B, u_3^B) = (0.996,\ 0.982,\ 0.987)$,
  $(u_1^C, u_2^C, u_3^C) = (0.989,\ 0.996,\ 0{,}995)$.

1081	*1082*	*1083*	*1084*	*1085*	*1086*	*1087*	*1088*	*1089*	*1090*
606.86	607.14	620.42	635.17	636.85	647.00	659.07	659.57	682.69	689.88

1091	*1092*	*1093*	*1094*	*1095*	*1096*	*1097*	*1098*	*1099*	*1100*
693.56	697.14	705.57	***A***	***B***	***C***	771.12	778.60	799.10	845.77

a) Wyznaczyć różnicę między łącznymi wymogami kapitałowymi dla analizowanych linii biznesu wyznaczonymi metodą **M1** i **M2**.
b) Porównać zastosowane metody agregacji wymogów kapitałowych, wskazując ich zalety i wady.

Odpowiedzi:

.....

Odp. a)
Różnica miedzy wymogami $W - \kappa = 205.87$.

.....

Odp. b)

W odpowiedzi należało:

- Zauważyć, że nie można było porównać wymogów kapitałowych wyznaczonych metodą M1 i M2. W przypadku M1 agregowano wartości zagrożone (VaR-y), w M2 - kapitały ekonomiczne wyznaczone jako $VaR(X) - E(X)$.
  Uwaga! Akceptowane było rozwiązanie, w którym w metodzie M1 agregowano kapitały ekonomiczne, czyli $VaR(X) - E(X)$. Wtedy $W - \kappa = -5.94$.
- Zwrócić uwagę dla jakich struktur zależności (między agregowanymi ryzkami) można stosować obydwie metody.
- Zwrócić uwagę na to, która z metod jest łatwiejsza, a która trudniejsza w praktycznym wykorzystaniu.

Rozwiązanie:

Ad a)

Metoda **M1**
Wymóg dla LOB1: $VaR_{0.995}(X_1) = F_1^{-1}(0.995) = 371.51$

Wymóg dla LOB2:
Dystrybuanta dla rozkładu wykładniczego ma postać $F_2(x) = 1 - e^{-\lambda x}$, stąd kwantyl rzędy α wyznaczamy z równania $\alpha = 1 - e^{-\lambda x}$. Otrzymujemy:

$$x_\alpha = \frac{\ln(1-\alpha)}{-\lambda}.$$

Zatem $VaR_{0.995}(X_2) = \frac{\ln(1-0.995)}{-0.02} = 264.92$

Wymóg dla LOB2: $VaR_{0.995}(X_3) = 100 + 20 \cdot \Phi^{-1}(0.995) = 151.52$

Łączny wymóg:

$$W = \sqrt{[371.51 \quad 264.92 \quad 151.52] \cdot \begin{bmatrix} 1 & 0.6 & 0.5 \\ 0.6 & 1 & 0.8 \\ 0.5 & 0.8 & 1 \end{bmatrix} \cdot \begin{bmatrix} 371.51 \\ 264.92 \\ 151.52 \end{bmatrix}} = 685.40$$

Metoda **M2**
W celu oszacowania VaR-u korzystamy ze wzoru:

$$\widehat{VaR}_\alpha(X) = X_{[n\cdot\alpha]+1:n}\,.$$

$\widehat{VaR}_{0.995}(S) = S_{[1100\cdot 0.995]+1:1100} = S_{1095:1100}$, czyli należało wyznaczyć wartość sumy dla ***B*.**

$$\boldsymbol{B} = F_1^{-1}(0.996) + F_2^{-1}(\,0.982) + F_2^{-1}(\,0.987) = 384.14 + 200.87 + 144.52 = 729.53$$

Łączny wymóg

$$\kappa = VaR_{0.995}(S) - E(S) = 729.53 - (100 + 50 + 100) = 479.53$$

Różnica: $W - \kappa = 205.87$

Ad. b) Patrz odpowiedź do b).

Sesja egzaminacyjna w dniu 19 listopada 2019 r.

Modelowanie

Arkusz ocen

Zadanie nr	Punktacja
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	