Komisja Egzaminacyjna dla Aktuariuszy

LXXXIX Egzamin dla Aktuariuszy

Sesja egzaminacyjna w dniu 27 lutego 2024 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 modeli:
- 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,
- 'log Lik.' – logarytm wiarygodności.

c) 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

d) 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

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

Zadanie 1.

a) ***(2p.)*** Podaj definicje dewiancji i skalowanej dewiancji (*scaled deviance*). Wskaż, której wielkości, związanej z jakością modelu regresji liniowej, odpowiada dewiancja. Jaki rozkład ma skalowana dewiancja?

b) ***(1p.)*** Twoim zadaniem jest wybór modelu o najlepszych zdolnościach predykcyjnych spośród zagnieżdżonych uogólnionych modeli liniowych. Wyjaśnij dlaczego podczas wyboru takiego modelu nie tylko dewiancja powinna zostać uwzględniona.

c) ***(2p.)*** Oszacowano dwa zagnieżdżone uogólnione modele liniowe: *M1* i *M2*. W modelu *M2* w porównaniu z *M1* uwzględniono dodatkowo jakościową zmienną objaśniającą Z, przyjmującą wartości ze zbioru liczącego 6 kategorii. Zbiór uczący liczył 3000 obserwacji. Uzyskano między innymi następujące informacje:

	Model nasycony (*Saturated Model*)	Model M1	Model *M2*
Logarytm wiarygodności (*Log-Likelihood*)	-700	-990	-985
Oszacowanie parametru dyspersji	2.0	2.0	2.0

Wykorzystując odpowiedni test (na poziomie istotności 0.05) i wybrane kryterium informacyjne, sprawdź, czy uwzględnienie zmiennej Z poprawiło zdolności predykcyjne modelu *M2* (w porównaniu z *M1*).

Odpowiedzi:

.....

Odp. a)

Zobacz podrozdział 4.5.3 w: "Effective Statistical Learning Methods for Actuaries I" - M. Denuit, D. Hainaut, J. Trufin, Springer, 2019.

.....

Odp. b)

Dodanie kolejnej zmiennej do modelu wpływa na obniżenie dewiancji. Jednak wykorzystanie tylko dewiancji może prowadzić do budowy modelu z nadmierną liczbą parametrów oraz do jego przeuczenia.

.....

Odp. c)

Nie poprawiło.

Rozwiązanie:

Korzystamy ze statystyki F:

$$F = \frac{D\left(y; \hat{\theta}^{M1}\right) - D\left(y; \hat{\theta}^{M2}\right)}{\hat{\phi}(q-p)},$$

gdzie q i p oznaczają liczbę parametrów odpowiednio modelu M1 i M2.

Obliczenia:

$$D\left(y;\hat{\theta}^{M1}\right) - D\left(y;\hat{\theta}^{M2}\right) = 2(\ln(L_{M2}) - \ln(L_{M1})) = 2\left(-985 - (-990)\right) = 10$$

ponieważ:

$D\left(y;\hat{\theta}^{M1}\right) = 2(\ln(L_{Sat}) - \ln(L_{M1}))$,

$D\left(y;\hat{\theta}^{M2}\right) = 2(\ln(L_{Sat}) - \ln(L_{M2}))$,

gdzie L_{Sat} wiarygodność modelu nasyconego.

Stąd:

$$F = \frac{10}{2 \cdot 5} = 1.$$

Wartość krytyczna:

$$F_{kr} = F_{5;\infty} \approx 2.216$$

Kryteria informacyjne:

$$AIC_A = -2\ln(L_{M1}) + 2p = 2 \cdot 990 + 2p = 1980 + 2p$$

$$AIC_B = -2\ln(L_{M2}) + 2(p+5) = 2 \cdot 985 + 2p + 10 = 1980 + 2p$$

Czyli:

$$AIC_{M1} = AIC_{M2}$$

Zadanie 2.

a) ***(2p.)*** Co to jest „*uporządkowana*” krzywa Lorenza (*ordered Lorenz curve*)? W jaki sposób może być wykorzystana w taryfikacji?

b) ***(3p.)*** W procesie taryfikacji wykorzystywany jest model P1. Aktuariusz opracował inny konkurencyjny model P2. Modele te przetestował na zbiorze liczącym 5 obserwacji. Wyniki, które uzyskał przedstawia tabela 2.1. Skonstruuj „*uporządkowaną*” krzywą Lorenza. Krzywą przedstaw na rysunku 2.1 (w części Odp. b)). Opisz osie.

Tab. 2.1

Ryzyko	Składki		Szkody
	Model P2	Model P1	
1	4,5	5	4
2	7	8	6
3	5,5	6	5
4	2	2	5
5	6	4	5

Odpowiedzi

.....

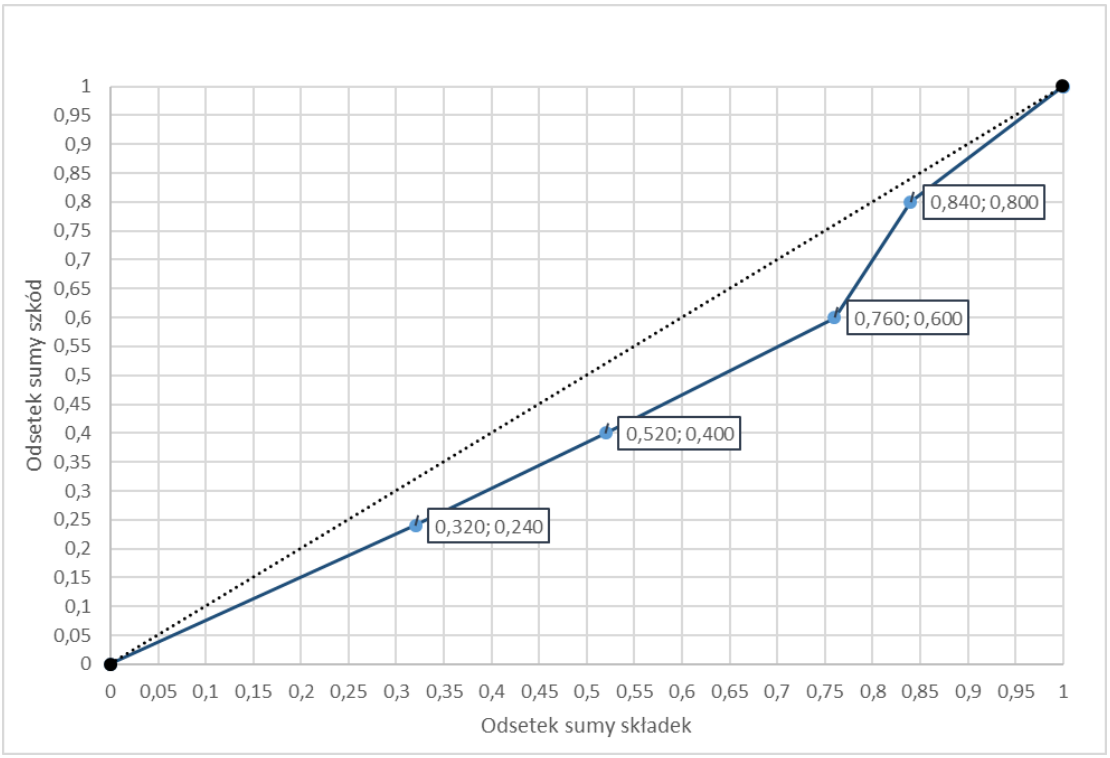
Odp. a)

Zobacz podrozdział 6.3.7 w: “Effective Statistical Learning Methods for Actuaries II” - M. Denuit, D. Hainaut, J. Trufin, Springer, 2020.

.....

Odp. b)

Rys. 2.1



Rozwiązanie:
Obliczenia pomocnicze:

Model P2 $(\hat{\mu}_2)$	Model P1 $(\hat{\mu}_1)$	Szkody L	$R_i = \frac{\hat{\mu}_2}{\hat{\mu}_1}$	Skumulo. $\hat{\mu}_1$	Skumulo. L	Odsetek sumy składek $\hat{\mu}_1$	Odsetek sumy szkód L
7	8	6	0,88	8	6	0,320	0,240
4,5	5	4	0,90	13	10	0,520	0,400
5,5	6	5	0,92	19	15	0,760	0,600
2	2	5	1,00	21	20	0,840	0,800
6	4	5	1,50	25	25	1,000	1,000

Uwaga! Obserwacje porządkujemy według R_i.

Zadanie 3.

Liczbę szkód (zmienna *L.szkod*) w pewnym portfelu ubezpieczeń AC modelowano z uwzględnieniem dwóch następujących zmiennych objaśniających:
W.Kierowcy – wiek kierowcy (zmienna jakościowa przyjmująca kategorie: W1, W2, W3),
W.Samoch – wiek samochodu (zmienna jakościowa przyjmująca kategorie: S1, S2).
Zebrano dane dotyczące liczby szkód zgłoszonych przez 67465 kierowców i przedstawiono je w tabeli 3.1 (w nawiasach podano ekspozycję na ryzyko).

Tab. 3.1

		W.Kierowcy		
		W1	W2	W3
W.Samoch	S1	736 (3677.36)	1315 (8312.54)	179 (1271.49)
	S2	788 (4825.76)	1706 (11882.75)	211 (1828.04)

a) ***(3p.)*** Twoim zadaniem jest oszacowanie, w oparciu o te dane, modelu regresji Poissona (z linkiem kanonicznym), z uwzględnieniem obydwu zmiennych objaśniających. Zakoduj zmienne objaśniające, przyjmując jako kategorie referencyjne: W1 i S1. Podaj:
   – postać macierzy modelu (*model matrix*),
   – wektor zmiennej zależnej (obserwacji)
   – wektor przedstawiający ekspozycję na ryzyko.

c) ***(2p.)*** Po oszacowaniu modelu dysponujesz następującymi wynikami:
   – oszacowania parametrów

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.64303    0.02982 -55.106  < 2e-16 ***
W.KierowcyW2 -0.17786    0.03143  -5.660 1.52e-08 ***
W.KierowcyW3 -0.35066    0.05675  -6.179 6.45e-10 ***
W.SamochS2   -0.13818    0.02861  -4.830 1.36e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 79.1439  on 5  degrees of freedom
Residual deviance:  3.2361  on 2  degrees of freedom
AIC: 60.699
```

   – analiza dewiancji

```
           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                          5     79.144
W.Kierowcy  2   52.698         3     26.446 3.604e-12 ***
W.Samoch    1      ?          2      3.236    ?
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Czy model z obydwoma zmiennymi objaśniającymi ma istotnie mniejszą dewiancję w porównaniu z modelem, w którym jedyną zmienną objaśniającą jest wiek kierowcy (*W.Kierowcy*)? Odpowiedź uzasadnij powołując się na wyniki odpowiedniego testu na poziomie istotności 0.05.

Odpowiedzi

.....

Odp. a)

- Macierz modelu (*model matrix*):

(Intercept)	*W.KierowcyW2*	*W.KierowcyW3*	*W.SamochS2*
1	0	0	0
1	1	0	0
1	0	1	0
1	0	0	1
1	1	0	1
1	0	1	1

- Wektor zmiennej zależnej (obserwacji)

L.szkod
736
1315
179
788
1706
211

- Wektor przedstawiający ekspozycję na ryzyko.

Ekspozycja
3677.36
8312.54
1271.49
4825.76
11882.75
1828.04

Szczegóły, zobacz podrozdział 4.2 w: "Effective Statistical Learning Methods for Actuaries I" - M. Denuit, D. Hainaut, J. Trufin, Springer, 2019.

.....

Odp. b)

Tak, model z obydwoma zmiennymi objaśniającymi ma istotnie mniejszą dewiancję w porównaniu z modelem, w którym jedyną zmienną objaśniającą jest wiek kierowcy (*W.Kierowcy*).

Rozwiązanie:

W zadaniu należy porównać dwa modele zagnieżdżone, gdy parametr dyspersji (skali) jest znany. W związku z tym można zastosować test chi-kwadrat dla różnicy dewiancji. Statystyka testowa wyraża się wzorem:

$$\chi^2 = D\left(y; \hat{\theta}^P\right) - D\left(y; \hat{\theta}^Q\right),$$

gdzie:

$D(y;\hat{\theta}^P)$ – dewiancja modelu o mniejszej liczbie parametrów p,
$D(y;\hat{\theta}^Q)$ – dewiancja modelu o większej liczbie parametrów q,
Statystka ta ma rozkład chi-kwadrat o $q-p$ stopniach swobody
Wartość statytstyki:

$$\chi^2 = D(y;\hat{\theta}^P) - D(y;\hat{\theta}^Q) = 26.446 - 3.236 = 23.21$$

Stopnie swobody: $3-2=1$
Wartość krytyczna na poziomie istotności 0.05 wynosi (z tablic): 3.841.

Wniosek: Na poziomie istotności 0.05, model z obydwoma zmiennymi objaśniającymi ma istotnie mniejszą dewiancję w porównaniu z modelem, w którym jedyną zmienną objaśniającą jest wiek kierowcy (*W.Kierowcy*).

Zadanie 4.
Obserwowano 100 000 osób od dokładnie 50-tego roku życia przez 30 lat. Zebrane dane zgrupowano i podano w następującej tabeli (tab. 4.1):

Tab. 4.1

Wiek	Liczba zgonów
[50, 60)	1700
[60, 70)	4700
[70 – 75)	5600
[75 – 80]	9700

a) (**2*p.***) Na podstawie danych z tab. 4.1 wyznacz krzywą ogiwalną dla funkcji przeżycia (*ogive empirical survival function*).
b) (**3*p.***) Wykorzystując skonstruowaną krzywą oblicz prawdopodobieństwo przeżycia $\hat{S}_{50}(t)$ dla $t = 2, t = 17$ i $t = 28$.

Odpowiedzi

.....

Odp. a)

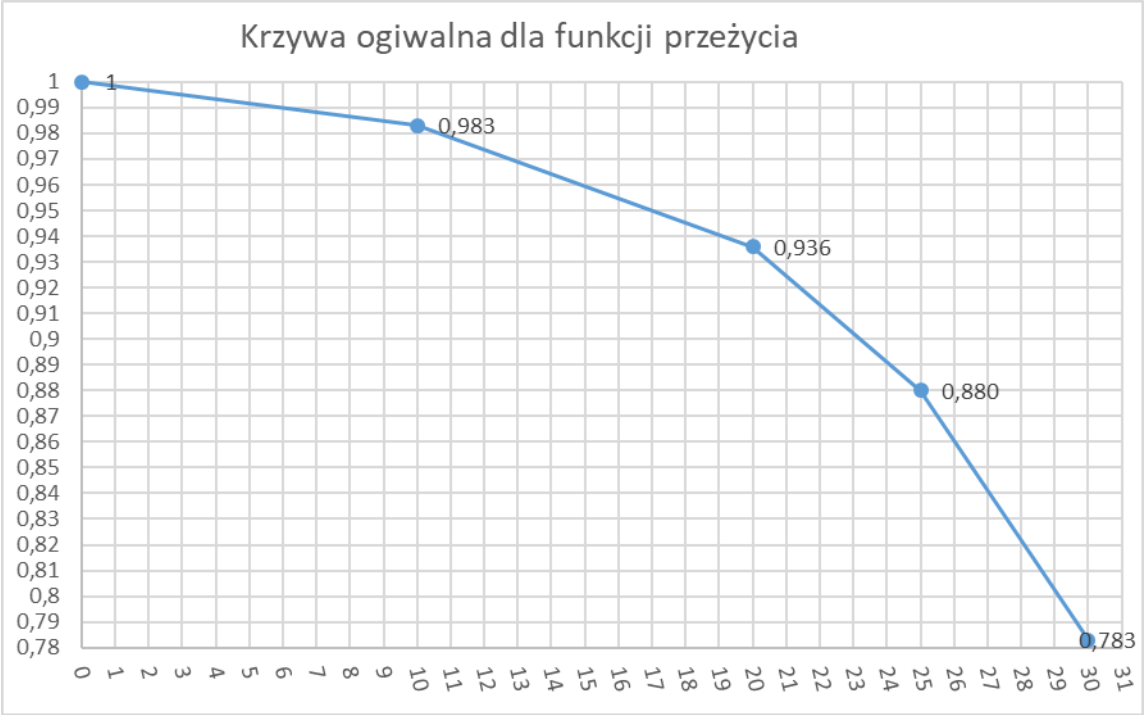
$$\hat{S}_{50}(0) = 1$$

$$\hat{S}_{50}(10) = \frac{100000 - 1700}{100000} = 0.983$$

$$\hat{S}_{50}(20) = \frac{100000 - (1700 + 4700)}{100000} = 0.936$$

$$\hat{S}_{50}(25) = \frac{100000 - (1700 + 4700 + 5600)}{100000} = 0.880$$

$$\hat{S}_{50}(30) = \frac{100000 - (1700 + 4700 + 5600 + 9700)}{100000} = 0.783$$



Szczegóły, zobacz podrozdział 18.4 w: "Actuarial Mathematics for Life Contingent Risks", 3rd edition - D. Dickson, M. Hardy, H. Waters, Cambridge, 2020.

.....

Odp. b)

$$\hat{S}_{50}(2) = \frac{8 \cdot 1 + 2 \cdot 0.983}{10} = 0{,}9966$$

$$\hat{S}_{50}(17) = \frac{3 \cdot 0.983 + 7 \cdot 0.936}{10} = 0{,}9501$$

$$\hat{S}_{50}(28) = \frac{2 \cdot 0.88 + 3 \cdot 0.783}{5} = 0{,}8218$$

Szczegóły, zobacz podrozdział 18.4 w: "Actuarial Mathematics for Life Contingent Risks", 3rd edition - D. Dickson, M. Hardy, H. Waters, Cambridge, 2020.

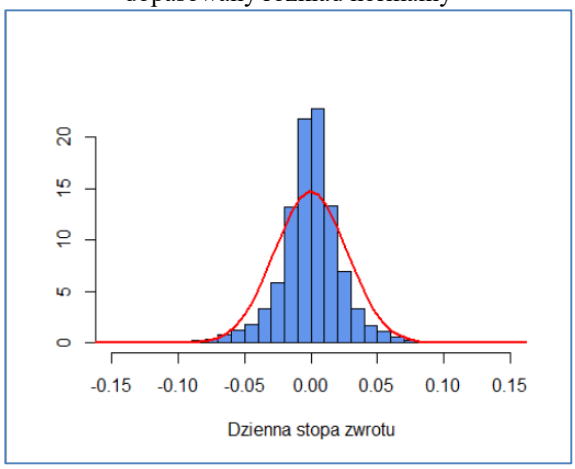
Zadanie 5.

a) (**3*p.***) Wymień i krótko scharakteryzuj trzy wybrane właściwości finansowych szeregów czasowych (tzw. stylizowane fakty).

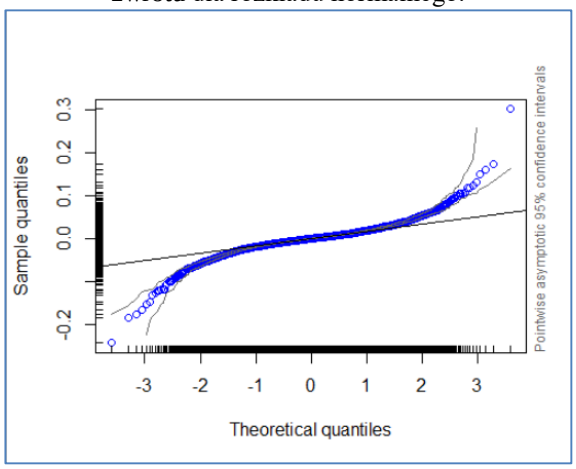
b) (**2*p.***) Na rysunku 5.1 przedstawiono 4 wykresy ilustrujące różne właściwości finansowego szeregu czasowego dziennych logarytmicznych stóp zwrotu spółki Aegon. Spośród wykresów A, B, C i D wybierz co najmniej trzy i przypisz im odpowiednią własność. (Uwaga, **dwa** punkty można uzyskać, gdy poprawnie zostaną zidentyfikowane co najmniej trzy stylizowane fakty, **jeden** - co najmniej dwa.)

Rys. 5.1

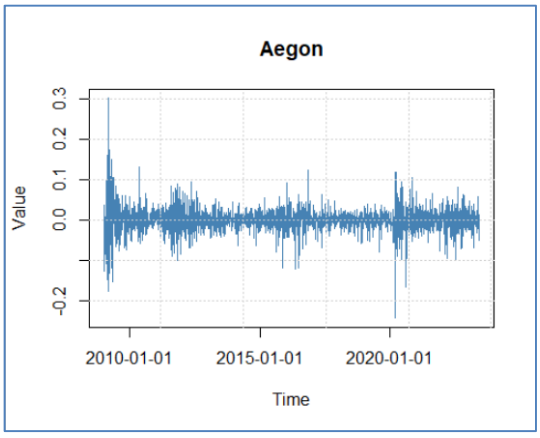
A. Histogram dla logarytmicznych stóp zwrotu i dopasowany rozkład normalny



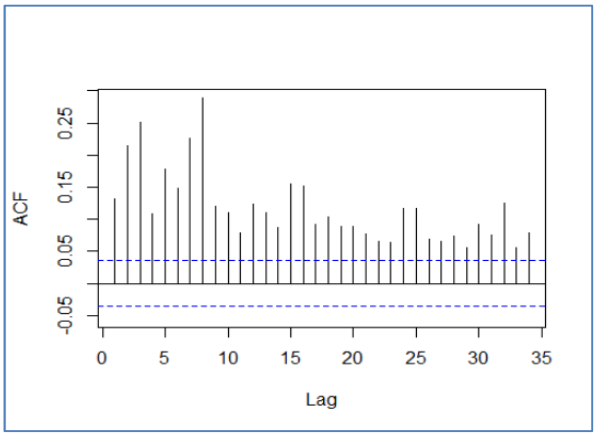
B. Wykres kwantyl-kwantyl logarytmicznych stóp zwrotu dla rozkładu normalnego.



C. Logarytmiczne stopy zwrotu



D. Autokorelacja modułów logarytmicznych stóp zwrotu



Odpowiedzi

.....

Odp. a)

Zobacz podrozdział 3.1 w: "Quantitative Risk Management: Concepts, Techniques and Tools", revised edition - A. McNeil, R. Frey, P. Embrechst, Princeton, 2015

.....

Odp. b)

Rysunek:	Własność
A	Nie podlegają rozkładowi normalnemu
B	Grube ogony
C	Grupowanie zmienności
D	Dodatnia autokorelacja modułów stóp zwrotu

Zadanie 6.

a) (**3*p.***) Na czym polega różnica między obserwacjami uciętymi (*truncated data*) a obserwacjami cenzurowanymi (*censored data*)? Wskaż i omów co najmniej jedną sytuację, w której aktuariusz może wykorzystywać:
   i. obserwacje ucięte,
   ii. obserwacje cenzurowane.

b) (**2*p.***) Wiadomo, że szkody w pewnym portfelu ubezpieczeń mają rozkład Weibulla z parametrem $\tau = 2$. W portfelu tym odnotowano 5 szkód. Wysokości trzech z nich były równe: 20, 30 i 45 tys. zł. O dwóch pozostałych wiadomo, że przekroczyły 50 tys. zł. Metodą największej wiarygodności oszacuj parametr θ tego rozkładu.

Uwaga! Dla rozkładu Weibulla:

$$f(x) = \frac{\tau \cdot \left(\frac{x}{\theta}\right)^{\tau} \cdot e^{-\left(\frac{x}{\theta}\right)^{\tau}}}{x}, \qquad F(x) = 1 - e^{-\left(\frac{x}{\theta}\right)^{\tau}}$$

Odpowiedzi:

.....

Odp. a)

Zobacz podrozdziały: 11.4, 14.3 i 14.5 w: "Loss Models: From Data to Decisions", 5th edition - S.A. Klugman, H.H Panjer, G.E. Willmot, Wiley, 2019.

.....

Odp. b)

$\hat{\theta} = 52.68$

Rozwiązanie:

Funkcja wiarygodności:

$$L(\theta) = f(20) \cdot f(30) \cdot f(45) \cdot \left(1 - F(50)\right)^2$$
$$\propto \theta^{-2} e^{-\left(\frac{20}{\theta}\right)^2} \theta^{-2} e^{-\left(\frac{30}{\theta}\right)^2} \theta^{-2} e^{-\left(\frac{45}{\theta}\right)^2} \left(e^{-\left(\frac{50}{\theta}\right)^2}\right)^2$$
$$= \theta^{-6} e^{-8325/\theta^2}$$

Logarytm wiarygodności:

$$l(\theta) = -6 \ln \theta - 8325\theta^{-2}$$
$$l'^{(\theta)} = -6\theta^{-1} + 16650\theta^{-3}$$

Stąd:

$$\hat{\theta} = 52.68$$

Zadanie 7.
Jako aktuariusz wykorzystujesz modele bazujące na drzewach decyzyjnych. Chcesz skonstruować model o jak najlepszych zdolnościach predykcyjnych. Rozważasz możliwość wykorzystania metody uczenia zespołowego *boosting*.

a) (**2*p.***) Krótko opisz na czym polega ta metoda. Czy metoda *boosting* może być stosowana zarówno dla problemów regresji, jak i klasyfikacji?
b) (**1*p.***) W jaki sposób metoda *boosting* różni się od metody *bagging* pod względem sposobu wykorzystania danych treningowych?
c) (**2*p.***) Wymień i krótko opisz co najmniej dwa parametry dostrajania w metodzie *boosting*.

Odpowiedzi:

.....

Odp. a)
Zobacz np. podrozdział 8.2 w "An Introduction to Statistical Learning with Applications in R" - G. James, D. Witten, T. Hastie, R. Tibshirani, Springer, 2021

.....

Odp. b)
Zobacz np. podrozdział 8.2.3 w "An Introduction to Statistical Learning with Applications in R" - G. James, D. Witten, T. Hastie, R. Tibshirani, Springer, 2021

.....

Odp. c)
Zobacz np. podrozdział 8.2.3 w "An Introduction to Statistical Learning with Applications in R" - G. James, D. Witten, T. Hastie, R. Tibshirani, Springer, 2021

Zadanie 8.

a) (**2*p.***) Krótko omów współczynnik zależności V Cramera. Wskaż jakie może przyjmować wartości i co one oznaczają w kontekście analizy siły zależności między zmiennymi. W jakich przypadkach można stosować ten współczynnik? Czy jest on ograniczony tylko do zmiennych jakościowych, czy może być używany także dla zmiennych ilościowych? Jeżeli tak, to w jaki sposób można go obliczyć?

b) (**3*p.***) Liczbę szkód w pewnym portfelu ubezpieczeń AC modelowano z uwzględnieniem dwóch następujących zmiennych objaśniających:
Plec – płeć kierowcy (zmienna jakościowa: *K* (kobieta), *M* (mężczyzna)),
Dystans – przebyty dystans w ciągu roku (zmienna jakościowa, przyjmująca dwie kategorie: *poniżej 20 tys. km*, *powyżej. 20 tys. km*).
Zebrano dane dotyczące liczby szkód zgłoszonych przez 3000 kierowców i przedstawiono je w tabeli 8.1 (w nawiasach podano ekspozycję na ryzyko):

Tab. 8.1

		Dystans	
		poniżej 20 tys. km	*powyżej. 20 tys. km*
Plec	*K*	15 (200)	197 (1800)
	M	28 (600)	35 (400)

Wypowiedz się na temat zależności między zmiennymi objaśniającymi (tzn. między *Plec* a *Dystans*). Skorzystaj w tym celu ze współczynnika V Cramera. Czy ta zależność jest istotna statystycznie? Wykorzystaj odpowiedni test na poziomie istotności 0.05.

Uwaga! Wzór na współczynnik V Cramera:

$$V = \sqrt{\frac{\chi^2}{n \cdot \min\{k_1 - 1, k_2 - 1\}}}$$

Odpowiedzi:

.....

Odp. a)

Zobacz podrozdział 4.4.5 w: "Effective Statistical Learning Methods for Actuaries I" - M. Denuit, D. Hainaut, J. Trufin, Springer, 2019.

.....

Odp. b)

Występuje stosunkowo silna zależność między zmiennymi *Plec* a *Dystans* (współczynnik V Cramera wynosi 0.533). Na podstawie testu niezależności CHI-Kwadrat można stwierdzić, że zależność ta jest statystycznie istotna.

Rozwiązanie:
Statystyka χ^2 wyraża się wzorem:

$$\chi^2 = \sum_{i=1}^{k_1}\sum_{j=1}^{k_1}\frac{\left(n_{ij} - \hat{n}_{ij}\right)^2}{\hat{n}_{ij}}$$

gdzie:
n_{ij} – liczebności zaobserwowane,
$\hat{n}_{ij} = \frac{n_{i\cdot}\cdot n_{\cdot j}}{n}$ – liczebności teoretyczne (w tabeli na żółtym tle),
$n_{i\cdot}, n_{\cdot j}$ – liczebności brzegowe,
n – liczba obserwacji,
k_1, k_1 – odpowiednio liczba wierszy i kolumn w tablicy kontyngencji.

Statystyka χ^2 ma rozkład Chi-Kwadrat o $(k_1 - 1)(k_2 - 1)$ stopniach swobody.

Obliczenia pomocnicze:

	poniżej 20 tys. km	*powyżej. 20 tys. km*	Suma
M	533.333 200	1466.667 1800	2000
K	266.667 600	733.333 400	1000
Suma	800	2200	3000

$$\chi^2 = 852.273$$

Wartość krytyczna na poziomie istotności 0.05 wynosi 3.841. Czyli występuje istotny statystycznie związek między zmiennymi *Plec* a *Dystans*.

Współczynnik V Cramera:

$$V = \sqrt{\frac{852.273}{3000 \cdot \min\{2-1,\ 2-1\}}} = 0.533$$

Zadanie 9.

Dysponujesz następującymi danymi dotyczącymi wysokości szkód w pewnym portfelu ubezpieczeń (tab. 9.1):

Tab. 9.1

Nr	Szkoda	Nr	Szkoda
1	6.163	11	1.484
2	5.618	12	1.392
3	5.542	13	1.313
4	4.037	14	1.170
5	1.869	15	1.148
6	1.795	16	1.112
7	1.722	17	1.107
8	1.579	18	1.061
9	1.570	19	1.043
10	1.533	20	1.006

Szkody są uporządkowane od największej do najmniejszej.

a) ***(1p.)*** Co to jest estymator Hilla i w jaki sposób jest używany do analizy danych?
b) (***2p.***) Na podstawie danych z tab. 9.1 skonstruowano wykres Hilla i przedstawiono go na rysunku 9.1 (część Odp. b)). Opisz osie wykresu i uzupełnij brakujący fragment.
c) (***2p.***) Zaproponowano, aby wysokości szkód w tym portfelu były modelowane z wykorzystaniem rozkładu o dystrybuancie: $F(x) = 1 - \left(\frac{1}{x}\right)^{1.9}, x > 1$. Czy uznajesz tę propozycję za słuszną? Odpowiedź uzasadnij!

Uwaga! Estymator Hilla ma postać:

$$\hat{\alpha}_{k,n} = \left(\frac{1}{k} \sum_{j=1}^{k} \ln X_{j,n} - \ln X_{k,n} \right)^{-1}, 2 \leq k \leq n$$

Odpowiedzi:

.....

Odp. a)

Zobacz podrozdział 5.2.4 w: "Quantitative Risk Management: Concepts, Techniques and Tools", revised edition - A. McNeil, R. Frey, P. Embrechst, Princeton, 2015

.....

Odp. b)

Obliczenia pomocnicze:

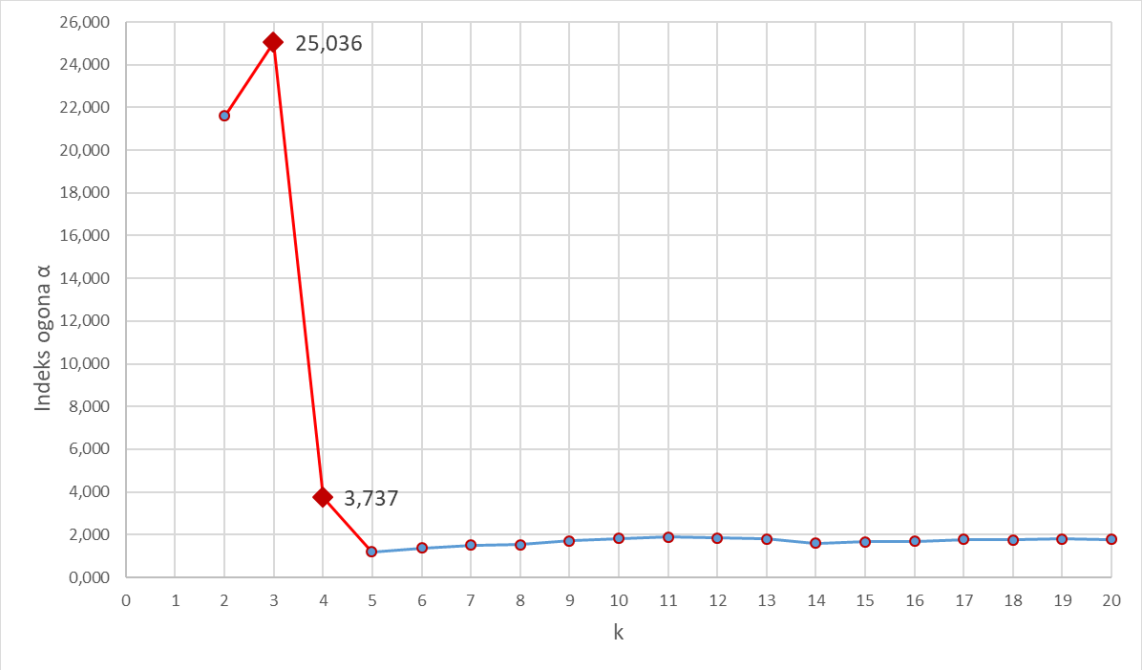
$$\hat{\alpha}_{3,n} = \left(\frac{1}{3}\sum_{j=1}^{3} \ln X_{j,n} - \ln X_{3,n}\right)^{-1} =$$

$$\left(\frac{1}{3}(\ln 6.163 + \ln 5.618 - 2 \cdot \ln 5.542)\right)^{-1} = 25.036$$

Analogicznie:

$$\hat{\alpha}_{4,n} = 3.737$$

Rys. 9.1



.....

Odp. c)

Tak.

Jest to dystrybuanta rozkładu Pareto $F(x) = 1 - \left(\frac{\theta}{x}\right)^{\alpha}, x > \theta$, w którym $\theta = 1$, $\alpha = 1.9$. Parametr $\alpha = 1.9$ można uzasadnić wykresem Hilla („stabilizuje się" mniej więcej na poziomie 1.9). Parametr $\theta = 1$, jest to najmniejsza zaobserwowana szkoda.

Zadanie 10.

a) (**2*p.***) Przedstaw indeks Giniego jako kryterium dobroci podziału (*goodness of split criterion*) w drzewach decyzyjnych.

b) (**3*p.***) Dysponujesz następującym (prostym) zbiorem danych dla zmiennych dychotomicznych (tab. 10.1):

Tab. 10.1

Zmienna zależna Y	Zmienne niezależne	
	X_1	X_2
1	Tak	B
1	Nie	A
0	Tak	B
0	Tak	A
1	Tak	A
0	Nie	B
0	Tak	B

Twoim zadaniem jest zbudowanie drzewa klasyfikacyjnego, w którym jako kryterium dobroci podziału wybrano indeks Giniego. Ustal, która zmienna powinna znajdować się w korzeniu drzewa (węźle głównym). Wiadomo, że w przypadku zmiennej X_1 średnia ważona odpowiednich indeksów Giniego wynosi 0.4857.

Uwaga! Indeks Giniego:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

Odpowiedzi:

.....

Odp. a)

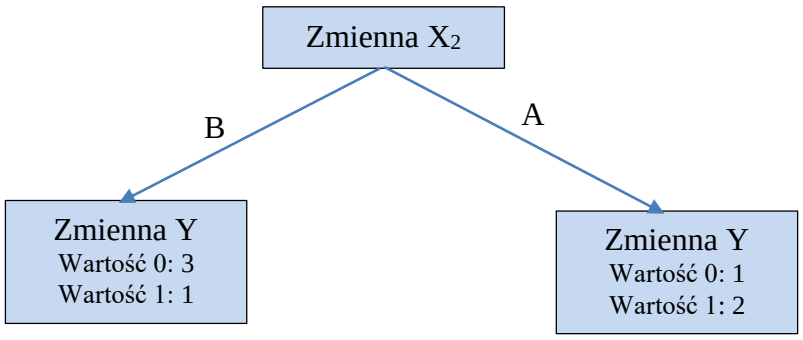
Zobacz np. podrozdział 8.1.2 w "An Introduction to Statistical Learning with Applications in R" - G. James, D. Witten, T. Hastie, R. Tibshirani, Springer, 2021

.....

Odp. b)

W korzeniu drzewa powinna znajdować się zmienna X_2 ($G_{sw} < 0.4857$).

Rozwiązanie:



Lewa gałąź: $G_L = \frac{1}{4} \cdot \frac{3}{4} + \frac{3}{4} \cdot \frac{1}{4} = 0.375$

Prawa gałąź: $G_L = \frac{1}{3} \cdot \frac{2}{3} + \frac{2}{3} \cdot \frac{1}{3} = 0.444$

Średnia ważona: $G_{sw} = \frac{4}{7} \cdot 0.375 + \frac{3}{7} \cdot 0.444 = 0.405$

Sesja egzaminacyjna w dniu 27 lutego 2024 r.

Modelowanie

Arkusz ocen

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