Komisja Egzaminacyjna dla Aktuariuszy

LXXXVI Egzamin dla Aktuariuszy

Sesja egzaminacyjna w dniu 20 września 2022 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) Wartości $\chi^2_{\alpha;\upsilon}$ rozkładu chi-kwadrat spełniające warunek $P(\chi^2 \geq \chi^2_{\alpha;\upsilon}) = \alpha$

υ \ α	***0.99***	***0.95***	***0.90***	***0.48***	***0.49***	***0.50***	***0.10***	***0.05***	***0.01***	***0.008***	***0.007***	***0.005***
1	0.000	0.004	0.016	0.499	0.477	0.455	2.706	3.841	6.635	7.033	7.273	7.879
2	0.020	0.103	0.211	1.468	1.427	1.386	4.605	5.991	9.210	9.657	9.924	10.597
3	0.115	0.352	0.584	2.474	2.420	2.366	6.251	7.815	11.345	11.827	12.115	12.838
4	0.297	0.711	1.064	3.486	3.421	3.357	7.779	9.488	13.277	13.789	14.094	14.860
5	0.554	1.145	1.610	4.499	4.425	4.351	9.236	11.070	15.086	15.625	15.946	16.750
6	0.872	1.635	2.204	5.512	5.430	5.348	10.645	12.592	16.812	17.375	17.710	18.548
7	1.239	2.167	2.833	6.525	6.435	6.346	12.017	14.067	18.475	19.060	19.408	20.278
8	1.646	2.733	3.490	7.537	7.440	7.344	13.362	15.507	20.090	20.696	21.056	21.955
9	2.088	3.325	4.168	8.548	8.445	8.343	14.684	16.919	21.666	22.291	22.663	23.589
10	2.558	3.940	4.865	9.559	9.450	9.342	15.987	18.307	23.209	23.853	24.235	25.188
11	3.053	4.575	5.578	10.570	10.455	10.341	17.275	19.675	24.725	25.386	25.779	26.757
12	3.571	5.226	6.304	11.580	11.460	11.340	18.549	21.026	26.217	26.895	27.297	28.300
13	4.107	5.892	7.042	12.589	12.464	12.340	19.812	22.362	27.688	28.383	28.794	29.819
14	4.660	6.571	7.790	13.599	13.469	13.339	21.064	23.685	29.141	29.851	30.272	31.319
15	5.229	7.261	8.547	14.608	14.473	14.339	22.307	24.996	30.578	31.303	31.732	32.801
16	5.812	7.962	9.312	15.617	15.477	15.338	23.542	26.296	32.000	32.740	33.178	34.267
17	6.408	8.672	10.085	16.626	16.481	16.338	24.769	27.587	33.409	34.162	34.609	35.718
18	7.015	9.390	10.865	17.634	17.485	17.338	25.989	28.869	34.805	35.573	36.027	37.156
19	7.633	10.117	11.651	18.642	18.489	18.338	27.204	30.144	36.191	36.972	37.434	38.582
20	8.260	10.851	12.443	19.650	19.493	19.337	28.412	31.410	37.566	38.360	38.830	39.997

d) 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.

W celu zbadania, w jaki sposób roczna liczba szkód (zmienna *ClaimNb*) dla polis w pewnym portfelu ubezpieczeń AC zależy od wieku kierowcy (zmienna *DriverAge* w latach), marki samochodu (zmienna *Brand* przyjmująca 7 kategorii) oraz rodzaju paliwa (zmienna *Gas* przyjmująca 2 kategorie) wykorzystano regresję Poissona z linkiem kanonicznym.

a) (**3*p.***) Najpierw oszacowano model ze wszystkimi wymienionymi wyżej zmiennymi objaśniającymi bez uwzględniania interakcji między nimi. Dla tego modelu tabela analizy dewiancji (*Deviance Table*) przedstawia się następująco:

Analysis of Deviance Table
Model: poisson, link: log
Response: *ClaimNb*
Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev
NULL			72854	22350
DriverAge	1	8.943	72853	22341
Brand	6	45.050	72847	22296
Gas	1	41.109	72846	22255

Wypowiedz się na temat istotności wpływu wybranych zmiennych na roczną liczbę szkód zgłaszanych w tym portfelu. Wykorzystaj w tym celu wyniki odpowiedniego testu (na poziomie istotności 0.05). Nazwij ten test i krótko opisz.

b) (**2*p.***) Powyższy model został rozszerzony o interakcję między marką samochodu a rodzajem paliwa. Kryterium informacyjne Akaike (AIC) dla tego modelu wynosi 29433.45. Wykorzystując odpowiedni test sprawdź, czy wszystkie współczynniki interakcji (łącznie) są statystycznie istotne. Przyjmij poziom istotności równy 0.05. W szczególności zapisz hipotezę zerową i alternatywną, podaj wartość statystyki testowej i jej rozkład oraz wartość krytyczną.
Wiadomo, że kryterium informacyjne Akaike (AIC) dla wyjściowego modelu (bez interakcji) wynosi 29425.22.

Odpowiedzi

.....

Odp. a)

Należało skorzystać z testu ilorazu wiarygodności.

- Dla zmiennej *DriverAge* wartość statystyki testowej wynosi 8.943, wartość krytyczna: $\chi^2_{0.05;1} = 3.841$. Możemy zatem przyjąć, że wiek kierowcy istotnie wpływa na roczną liczbę szkód zgłaszanych w tym portfelu.
- Dla zmiennej *Brand* wartość statystyki testowej wynosi 45.050, wartość krytyczna: $\chi^2_{0.05;6} = 12.592$. Możemy zatem przyjąć, że marka samochodu istotnie wpływa na roczną liczbę szkód zgłaszanych w tym portfelu.

- Dla zmiennej *Gas* wartość statystyki testowej wynosi 41.109, wartość krytyczna: $\chi^2_{0.05;1} = 3.841$. Możemy zatem przyjąć, że rodzaj paliwa istotnie wpływa na roczną liczbę szkód zgłaszanych w tym portfelu.

.....

Odp. b)

Współczynniki interakcji (łącznie) nie są statystycznie istotne.

Też należało skorzystać z testu ilorazu wiarygodności.

H_0: Wszystkie współczynniki interakcji między marką samochodu a rodzajem paliwa są równe zero.

H_1: Przynajmniej jeden współczynnik jest różny od zera.

Statystyka testowa:

$$T = 2(l_2 - l_1),$$

gdzie:

l_2, l_1 - logarytmy wiarygodności odpowiednio modelu z interakcjami i bez interakcji.

Wiadomo, że

$$AIC = -2 \cdot l + 2 \cdot k,$$

gdzie:

l - logarytm wiarygodności,

k - liczba parametrów.

Stąd:

$$T = -29433.45 + 2 \cdot 15 + 29425.22 - 2 \cdot 9 = 3.77$$

Wartość krytyczna: $\chi^2_{0.05;6} = 12.592$.

Rozwiązanie:

Zadanie 2.

Do opracowania planu taryfikacji wykorzystano uogólniony model liniowy. Uzyskano następujące wyniki:

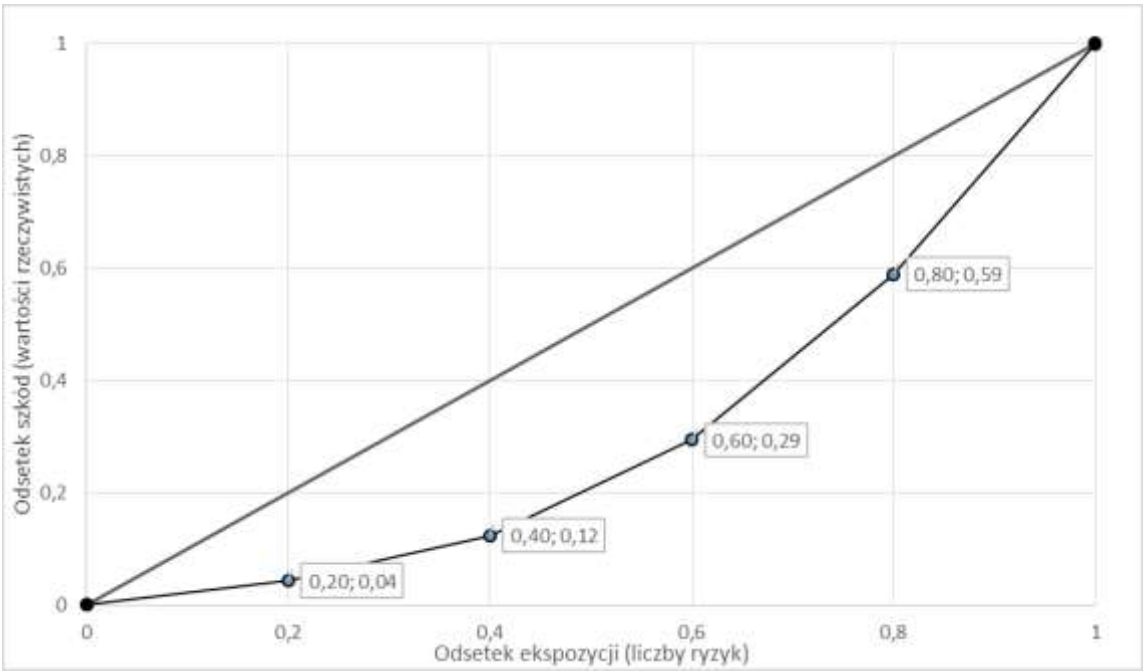
Ryzyko	Wartości prognozowane	Wartości rzeczywiste
1	2.00	2.05
2	0.50	0.22
3	1.50	1.48
4	0.80	0.85
5	0.80	0.40

a) **(2*p.*)** Narysuj krzywą Lorenza dla tego planu. Opisz osie i podaj współrzędne odpowiednich punktów na krzywej.
b) **(2*p.*)** Oblicz współczynnik Giniego dla tego planu.
c) **(1*p.*)** W jakim celu wykorzystuje się współczynnik Giniego w taryfikacji?

Odpowiedzi:

.....

Odp. a)



.....

Odp. b)

Współczynnik Giniego

$$G = 2 \cdot (0.5 - pole\ pod\ krzywą\ Lorenza)$$
$$G = 0.3792$$

.....

Odp. c)

Współczynnik Giniego zastosowany w taryfikacji mierzy poziom zróżnicowania składek w planach taryfowych. Im wyższa jego wartość, tym bardziej składki są zróżnicowane. Preferowane są plany taryfowe o wyższej wartości współczynnika Giniego.

Rozwiązanie:

Zadanie 3.

Do modelu regresji Poissona (z linkiem kanonicznym), który opisuje roczną liczbę szkód dla polis w pewnym portfelu ubezpieczeń, aktuariusz dodał nową ciągłą zmienną Z przyjmującą wartości od 1 do 100. Nowy model (regresji Poissona) uwzględniający Z oszacował, traktując wartości dopasowane wyjściowego modelu jako zmienną offsetową. Otrzymał następujące parametry:

	Estimate
(Intercept)	1.25
Z	-0.02

Wartości dopasowane wyjściowego modelu oraz wartości zmiennej Z dla trzech polis są następujące:

Polisa	Wartość dopasowana (model bez zmiennej Z)	Wartość zmiennej Z
1	0.013	45
2	0.019	92
3	0.025	35

a) (**3*p.***) Oblicz skorygowane prawdopodobieństwo wystąpienia co najmniej jednej szkody dla każdej z trzech powyższych polis.
b) (**2*p.***) Zidentyfikuj i krótko opisz sytuację, w której użycie offsetu jest lepsze niż (ponowne) dopasowanie wszystkich zmiennych.

Odpowiedzi:

.....

Odp. a)

Polisa 1: $\hat{\lambda} = 0.013 \cdot \exp(1.25 - 0.02 \cdot 45) = 0{,}018447878$
$P(K \geq 1) = 1 - P(K = 1) = 0{,}018278758$

Polisa 2: $\hat{\lambda} = 0.019 \cdot \exp(1.25 - 0.02 \cdot 92) = 0{,}010532218$
$P(K \geq 1) = 1 - P(K = 1) = 0{,}010476949$

Polisa 3: $\hat{\lambda} = 0.025 \cdot \exp(1.25 - 0.02 \cdot 35) = 0{,}043331325$
$P(K \geq 1) = 1 - P(K = 1) = 0{,}042405938$

.....

Odp. b)

Np.

- Jeżeli wprowadzając dodatkowe zmienne nie chcemy zmieniać już istniejących ze względu na pewne ograniczenia, jak np. systemu informatycznego, itp.
- W celu zrewidowania lub udoskonalenia istniejącego planu taryfowego (cennika). Jest on traktowany jako offset, a szacunkowe współczynniki regresji określają korekty potrzebne do jego poprawy.
- W celu ograniczania niektórych czynników ratingowych do przyjęcia wcześniej określonych wartości przy przechodzeniu z cennika technicznego na cennik

komercyjny. Takie ograniczenia są często motywowane konkurencją oraz względami prawnymi lub marketingowymi.

- **Inne przykłady w zalecanej literaturze.**

Rozwiązanie:

Zadanie 4.

a) (**2*p.***) Krótko omów model regresji Tweedie'ego.

b) (**1*p.***) Jaki jest główny problem z wykorzystaniem uogólnionego modelu liniowego ze zmienną zależną o rozkładzie złożonym Poissona (rozkładzie Tweedie z indeksem $1 < p < 2$) w taryfikacji? W jaki sposób można go rozwiązać?

c) (**2*p.***) W pewnym portfelu ubezpieczeń przeprowadzono taryfikację jednomodelową. W tym celu składkę czystą *Y* (*pure premium*) modelowano z wykorzystaniem regresji Tweedie'ego z linkiem logarytmicznym. Uwzględniono dwie następujące zmienne objaśniające:

- *wiek.kier* - wiek kierowcy. Zmienna jakościowa przyjmująca następujące kategorie: W1, W2, W3, W4, W5 i W6.
- *mar.sam* – marka samochodu. Zmienna jakościowa przyjmująca następujące kategorie: A, B, C, D i E.

Uzyskano następujące wyniki:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.57722	0.11597	39.469	< 2e-16 ***
*wiek.kier*W2	0.10033	0.15574	0.644	0.519450
*wiek.kier*W3	0.40314	0.17412	2.315	0.020601 *
*wiek.kier*W4	0.32675	0.15043	2.172	0.029856 *
*wiek.kier*W5	0.54284	0.15000	3.619	0.000296 ***
*wiek.kier*W6	0.84644	0.16220	5.219	1.81e-07 ***
*mar.sam*B	?	0.41871	-0.205	0.837761
*mar.sam*C	1.43077	0.19046	7.512	5.89e-14 ***
*mar.sam*D	0.46529	0.20745	2.243	0.024907 *
*mar.sam*E	0.03656	0.13064	0.280	0.779561

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Estimated dispersion parameter: 101.06

Estimated index parameter: 1.57

Residual deviance: 3626617 on 59990 degrees of freedom

AIC: 67379

Wykorzystując ten model, wyznaczy najwyższą ($\hat{Y}_{max}$) i najniższą ($\hat{Y}_{min}$) składkę czystą.

Odpowiedzi

.....

Odp. a)

Regresja Tweedie'ego to uogólniony model liniowy (GLM), w którym zmienna objaśniana ma rozkład Tweedie'ego, czyli rozkład należący do rodziny wykładniczej, dla którego funkcja wariancji ma postać: $\vartheta(\mu) = \mu^p$. Regresja, w której zmienna objaśnia ma rozkład Tweedie'ego z indeksem $1 < p < 2$ jest wykorzystywana w taryfikacji jednomodelowej.

.....

Odp. b)

Wykorzystując taki model w taryfikacji, musimy przyjąć, że czynniki ryzyka oddziałują w ten sam sposób zarówno na średnią liczbę szkód, jak i wysokość pojedynczej szkody. W praktyce nie zawsze mamy do czynienia z taką sytuacją. Przykładem mogą tu być ubezpieczenia komunikacyjne, w których geograficzne czynniki ryzyka nie rzadko mogą oddziaływać na liczę i wysokość szkód w przeciwnym kierunku. Rozwiązaniem może być zastosowanie podwójnego GLM.

(**Szczegóły w zalecanej literaturze.)**

.....

Odp. c)

Parametr przy zmiennej *mar.sam*B: $0.41871 \cdot (-0.205) = -0{,}085836$

$$\hat{Y}_{max} = 948{,}0715746$$
$$\hat{Y}_{min} = 89{,}24491551$$

Rozwiązanie:

Zadanie 5.

a) (**1*p.***) Zdefiniuj model autoregresyjny pierwszego rzędu (tj. AR(1))
b) (**1*p.***) Wskaż co najmniej dwa sposoby pozwalające stwierdzić, że model AR(1) może być odpowiednim kandydatem do modelowania danego szeregu czasowego.
c) (**1*p.***) Przedstaw sposób wyznaczania prognozy przedziałowej na okres $T+h, h=1,2,\ldots$ na podstawie modelu AR(1), T – długość szeregu czasowego (liczba obserwacji).
d) (**2*p.***) Dla pewnego szeregu czasowego liczącego 120 obserwacji ($T=120$) oszacowano następujący model:

ARIMA(1,0,0) with non-zero mean
Coefficients:

	ar1	mean
	0.7538	0.0495
s.e.	0.0594	0.1467

sigma^2 estimated as 0.1672: log likelihood=-62.38
AIC=130.76 AICc=130.96 BIC=139.12

Wartości szeregu dla 3 ostatnich okresów są równe:

t	*118*	*119*	*120*
y_t	0.363	0.434	0.328

Wykorzystując ten model, wyznacz prognozę przedziałową na poziomie ufności 0.95 na 122 okres.

Odpowiedzi:

.....

Odp. a)

Model autoregresyjny pierwszego rzędu AR(1) jest definiowany w następujący sposób:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t, \qquad t = 2, 3, \ldots, T,$$

gdzie: $\{\varepsilon_t\}$ jest procesem białego szumu, takim że $Cov(\varepsilon_{t+k}, y_t) = 0$ dla $k > 0$, a β_0, β_1 są nieznanymi parametrami.
Parametr β_0 może być dowolną stałą, natomiast β_1 może przyjmować wartości z przedziału od -1 do 1.

.....

Odp. b)

Na przykład:

- Diagram korelacyjny przedstawiający punkty $(y_t,\ y_{t-1})$. Powinny się układać wzdłuż linii prostej,
- Funkcja autokorelacji. Powinna znikać wykładniczo.
- Funkcja autokorelacji cząstkowej. Powinna być różna od zera dla $k = 1$.

.....

Odp. c)

Wykorzystując model AR(1) prognozę przedziałową z wyprzedzeniem h wyznaczana jest następująco:

$$\hat{y}_{T+h} \pm t_{1-\frac{\alpha}{2};T-3} \cdot \sqrt{s} \cdot \sqrt{1 + b_1^2 + \cdots + b_1^{2(h-1)}},$$

gdzie: $\hat{y}_{T+h}$ - prognoza punktowa z wyprzedzeniem h; $1-\alpha$ - poziom ufności (wiarygodnością prognozy); b_1 – oszacowanie parametru β_1; s – odchylenie stadardowe reszt; $t_{1-\frac{\alpha}{2};T-3}$ – kwantyl rzędu $1-\frac{\alpha}{2}$ rozkładu *t*-Studenta o $T-3$ stopniach swobody.

.....

Odp. d)

$$\hat{y}_{120+2} \pm t_{1-\frac{0.05}{2};117} \cdot \sqrt{0.1672} \cdot \sqrt{1 + 0.7538^2}$$

$$0.2077477 \pm 1.003619$$

Wartość $t_{1-\frac{0.05}{2};117}$ należało przybliżyć kwantylem rozkładu normalnego standaryzowanego $u_{1-\frac{0.05}{2};117}$, tj. wartością 1.96

Rozwiązanie:

Zadanie 6.

Poniżej podano dane dotyczące wysokości 225 roszczeń (w tys. zł.) dla pewnego portfela ubezpieczeń:

x_i	n_i
0 -10	100
10 - 20	40
20 - 40	30
40 - 70	28
70 -120	15
120 - 300	10
powyżej 300	2

a) (**1*p***) Przedstaw sposób wyznaczania funkcji gęstości empirycznej w oparciu o dane zgrupowane.
b) (**1*p***) Na podstawie podanych danych wyznacz funkcję gęstości empirycznej.
c) (**1*p***) Przedstaw sposób estymacji rozkładów metodą największej wiarygodności na podstawie danych zgrupowanych (podaj między innymi postać funkcji wiarygodności).
d) (**2*p***) Przyjęto, że roszczenia w analizowanym portfelu podlegają rozkładowi wykładniczemu. Wykorzystując podane dane wyznacz funkcję logarytmu wiarygodności dla parametru tego rozkładu.

Odpowiedzi:

.....

Odp. a)

Funkcja gęstości empirycznej dla danych zgrupowanych można wyznaczyć poprzez zróżniczkowanie ogiwalnej. Formuła jest następująca:

$$f_n(x) = \frac{n_j}{n(c_j - c_{j-1})}, \qquad c_{j-1} \le x < c_j$$

.....

Odp. b)

$$f_{225}(x) = \begin{cases} 0.044444, & 0 \le x < 10 \\ 0.017778, & 10 \le x < 20 \\ 0.006667, & 20 \le x < 40 \\ 0.004148, & 40 \le x < 70 \\ 0.001333, & 70 \le x < 120 \\ 0.000247, & 120 \le x < 300 \\ nieokreślona, & x \ge 300 \end{cases}$$

.....

Odp. c)

Funkcja wiarygodności:

$$L(\theta) = \prod_{j=1}^{k} \left(F(c_j|\theta) - F(c_{j-1}|\theta)\right)^{n_j}$$

Logarytm funkcji wiarygodności:

$$l(\theta) = \sum_{j=1}^{k} n_j \ln\left(F(c_j|\theta) - F(c_{j-1}|\theta)\right)$$

gdzie n_j oznacza liczbę obserwacji w przedziale $(c_{j-1},\ c_j]$.

.....

Odp. d)

Dystrybuanta rozkładu wykładniczego:

$$F(x) = \begin{cases} 1 - e^{-\theta x}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

Stąd

$l(\theta) = 100\ln(1 - e^{-10\theta}) + 40\ln(e^{-10\theta} - e^{-20\theta}) + 30\ln(e^{-20\theta} - e^{-40\theta}) + 28\ln(e^{-40\theta} - e^{-70\theta}) + 15\ln(e^{-70\theta} - e^{-120\theta}) + 10\ln(e^{-120\theta} - e^{-300\theta}) + 2\ln(e^{-300\theta})$.

Rozwiązanie:

Zadanie 7.

a) (**1*p.***) W jakim celu stosuje się techniki redukcji wariancji w metodzie Monte Carlo?
b) (**1*p.***) Wyznaczy błąd estymatora w klasycznej metodzie Monte Carlo.
c) (**3*p.***) Krótko opisz co najmniej dwie techniki redukcji wariancji.

Odpowiedzi:

.....

Odp. a)

W metodzie Monte Carlo jesteśmy zainteresowani minimalizacją błędu standardowego symulacji. Błąd ten można zmniejszyć zwiększając liczbę przebiegów symulacji (iteracji) lub stosując techniki redukcji wariancji. Pierwszy sposób jest mało efektywny. Drugi polega na redukcji odchylenia standardowego wartości zmiennej wynikowej. W tym celu można posłużyć się szeregiem metod, które modyfikują sposób próbkowania, tym samym przyczyniając się do bardziej regularnego pokrycia rozkładu prawdopodobieństwa analizowanego zjawiska. To z kolei skutkuje redukcją błędu.

.....

Odp. b)

Błąd wynosi

$$b = \frac{s}{\sqrt{n}},$$

gdzie

s – odchylenie standardowe wartości zmiennej wynikowej otrzymanych w poszczególnych iteracjach,

n – liczba iteracji.

.....

Odp. c)

Należało opisać dwie techniki redukcji wariancji, np. spośród niżej wymienionych:

- metodę zmiennych antytetycznych,
- losowanie warstwowe,
- metodę średniej ważonej,
- metodę zmiennych kontrolnych.

Szczegóły w zalecanej literaturze.

Zadanie 8.

a) (**2*p.***) Krótko opisz metodę estymacji jądrowej funkcji gęstości.
b) (**1*p.***) Jaką rolę w estymacji jądrowej odgrywa stała wygładzania?
c) (**2*p.***) Danych jest pięć obserwacji: 82, 126, 161, 294 i 384. Oszacuj wartość dystrybuanty $F(150)$, wykorzystując:
    a. estymację empiryczną,
    b. estymację jądrową z jądrem jednostajnym o stałej wygładzania 50.

Odpowiedzi:

.....

Odp. a)

Estymacja jądrowa polega na oszacowaniu nieznanej funkcji gęstości dla zmiennej losowej na podstawie skończonej liczby obserwacji tej zmiennej. Wartości funkcji gęstości w kolejnych punktach są wyznaczane jako względna częstość obserwacji w otoczeniu danego punktu. Otoczenie to nazwane jest pasmem estymacji (*bandwidth*). Do oszacowania względnej częstości wykorzystuje się funkcję gęstości zwaną funkcją jądra (*kernel*). Estymator jądrowy pozwala wyznaczyć funkcję gęstości bez konieczności uwzględniania z góry przyjętego rozkładu.

Szczegóły w zalecanej literaturze.

.....

Odp. b)

Stała wygładzania wpływa na zakres pasma estymacji, a tym samym na wygładzenie oszacowanej funkcji gęstości.

.....

Odp. c)

Estymacja empiryczna: $F(150) = \frac{2}{5} = 0.2$

Estymacja jądrowa:

Jądro jednostajne o stałej wygładzania 50:

$$k_y(x) = \begin{cases} 0, & x < y - 50 \\ \dfrac{1}{2 \cdot 50}, & y - 50 \le x \le y + 50 \\ 0, & x > y + 50 \end{cases}$$

Stąd otrzymujemy:

$$F(150) = 0.2 + \frac{74}{100} \cdot 0.2 + \frac{39}{100} \cdot 0.2 = 0.426$$

Rozwiązanie:

Zadanie 9.

a) (***2p.***) Przedstaw konstrukcję testu zgodności chi-kwadrat (podaj m. in. układ hipotez, postać statystyki testowej i jej rozkład, sposób ustalania wartości krytycznej).
b) (***2p.***) Z rozkładu jednostajnego na przedziale $(0, 1)$ wylosowano 1000 obserwacji. Wyniki pogrupowano w 20 przedziałów o takiej samej długości (tzn. przedstawiono w postaci danych zgrupowanych, w których przedziały $(c_{i-1}, c_i], i = 1,..,20$ mają taką samą długość). Następnie zsumowano kwadraty liczby obserwacji z poszczególnych klas i otrzymano wynik 51850 (tzn. $\sum_{i=1}^{20} n_i^2 = 51850$). W celu sprawdzenia zastosowanego generatora zastosowano test zgodności chi-kwadrat. Wyznacza prawdopodobieństwo testowe (p-value) dla tego testu.
c) (**1*p.***) Z rozkładu jednostajnego na przedziale $(0, 1)$ wylosowano 500 obserwacji i podobnie jak w punkcie b) pogrupowano je w 20 przedziałów o takiej samej długości. Okazało się, że liczebność każdego przedziału jest dwa razy mniejsza niż liczebność odpowiadającego mu przedziału z punktu b) (tzn. $n'_i = 0.5n_i, i = 1, \dots, 20$; n'_i, n_i – liczebności przedziałów odpowiednio z punktu c) i b), czyli uzyskano taki sam rozkład empiryczny jak w b)). Ile w tym przypadku wynosi prawdopodobieństwo testowe (p-value) w teście zgodności chi-kwadrat? Jaki stąd można wyciągnąć wniosek odnośnie wykorzystania tego testu?

Odpowiedzi:

.....

Odp. a)

Test zgodności chi-kwadrat służy do weryfikacji hipotezy, że rozkład badanej zmiennej ma dystrybuantę określonej postaci (przyjmuje określoną postać).
Układ hipotez:
$H_0: F(x) = F_0(x)$ - populacja generalna ma rozkład określony dystrybuantą $F_0(x)$,
$H_1: F(x) \neq F_0(x)$.
Wylosowaną z populacji próbę o liczebności *n* porządkujemy i tworzymy *k* rozłącznych klas. Przyjmując, że H_0 jest prawdziwa (tzn., że rozkład populacji generalnej opisany jest dystrybuantą $F_0(x)$) obliczamy prawdopodobieństwa p_j tego, że zmienna losowa przyjmuje wartości z j-tej klasy. Korzystając z prawdopodobieństw p_j obliczamy wartość statystyki

$$\chi^2 = \sum_{j=1}^{k} \frac{(n_j - np_j)^2}{np_j},$$

gdzie n_j jest liczebnością *j*-tej klasy.
Statystyka testowa ma asymptotyczny rozkład χ^2 o $k - q - 1$ stopniach swobody, gdzie q oznacza liczbę parametrów, które zostały oszacowane na podstawie rozkładu empirycznego.
Zbiór krytyczny: $[\chi^2_{\alpha;(k-q-1)}, +\infty)$, gdzie przy założonym poziomie istotności α wartość krytyczna $\chi^2_{\alpha;(k-q-1)}$ spełnia warunek $P(\chi^2 \geq \chi^2_{\alpha;(k-q-1)}) = \alpha$,

.....

Odp. b)

Wyznaczamy wartość statystyki:

$$\chi^2 = \sum_{j=1}^{20} \frac{(n_j - 50)^2}{50} = \frac{1}{50}\left(\sum_{j=1}^{20} n_j^2 - 100\sum_{j=1}^{20} n_j + 20 \cdot 50^2\right)$$
$$= \frac{1}{50}(51850 - 100 \cdot 1000 + 50000) = 37$$

Stąd prawdopodobieństwo testowe (p-value) wynosi (odczytane z zamieszczonych tablic rozkładu chi-kwadrat dla 19 stopni swobody): 0.008

.....

Odp. c)

$$\chi^2 = \sum_{j=1}^{k} \frac{(\frac{1}{2}n_j - \frac{1}{2}np_j)^2}{\frac{1}{2}np_j} = \frac{1}{2}\sum_{j=1}^{k} \frac{(n_j - np_j)^2}{np_j},$$

stąd $\chi^2 = 18.5$.

Prawdopodobieństwo testowe wynosi (odczytane z zamieszczonych tablic rozkładu chi-kwadrat dla 19 stopni swobody): 0.49

Należy być ostrożnym przy stosowaniu tego testu dla dużych *n*.

Rozwiązanie:

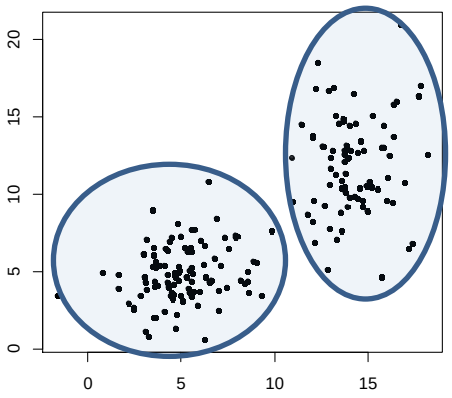
Zadanie 10.

a) (**2*p.***) Krótko opisz algorytm *k*-średnich.

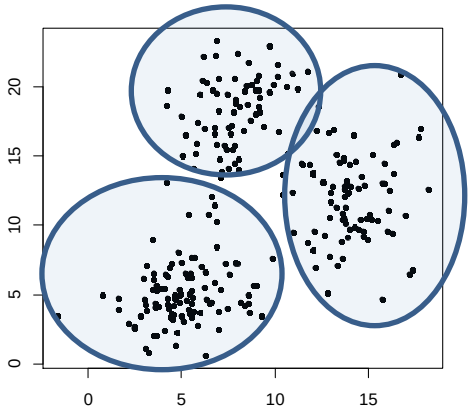
b) (**1*p.***) Podaj co najmniej dwa przykłady z dziedziny ubezpieczeń, gdzie można użyć metod grupowania (*cluster analysis*) i wskaż związane z tym korzyści.

c) (**2*p.***) W przypadku poniższych wykresów najpierw zdecyduj, ile jest skupień w każdym przypadku. Na podstawie wybranego *k* naszkicuj (na zamieszczonych wykresach) możliwy wynik algorytmu *k*-średnich. Dla którego wykresu algorytm nie zapewnia sensownego wyniku?
Nazwij właściwość, którą powinny mieć skupienia, aby algorytm dawał sensowne wyniki.

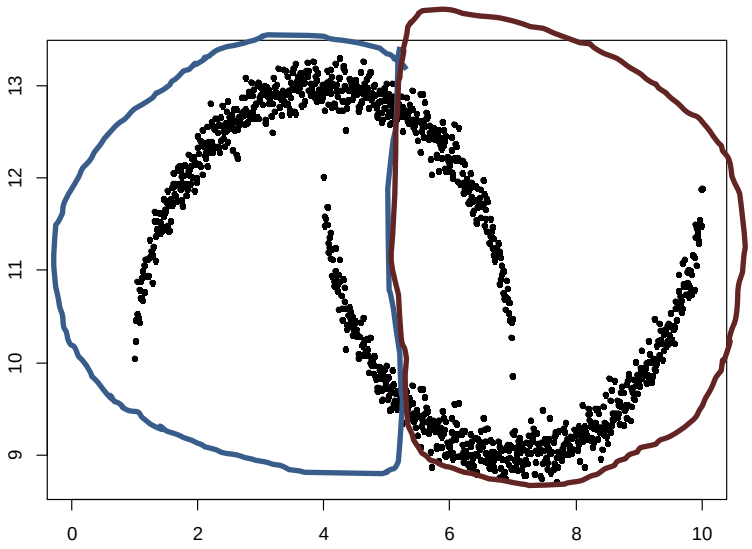
A



B



C



Odpowiedzi:

.....

Odp. a)

Punkty danych najpierw losowo przypisujemy do jednego z *k* klastrów. Z kolei, następujące kroki wykonujemy do momentu, gdy klastry przestaną się zmieniać:

- Obliczmy środki klastrów. Są to średnie arytmetyczne współrzędnych wszystkich punktów z danego klastra (są też inne techniki).
- Przypisujemy punkty danych do klastra, którego środek ma najmniejszą odległość.

Szczegóły w zalecanej literaturze.

.....

Odp. b)

Np.

- Metody grupowania można zastosować do segmentacji ubezpieczających. Segmentację można następnie wykorzystać w taryfikacji.
- W identyfikacji i analizie podejrzanych/kwestionowanych roszczeń ubezpieczeniowych.

.....

Odp. c)

Możliwy wynik algorytmu *k*-średnich przedstawiono na powyższych rysunkach. Oczywiście metoda ta nie daje sensownego wyniku w przypadku C. Aby metoda *k*-średnich dawała sensowne wyniki, poszczególne klastry powinny być w przybliżeniu okrągłe (lub sferyczne w wyższych wymiarach). Ta właściwość jest naruszona w trzecim przypadku (C).

Sesja egzaminacyjna w dniu 20 września 2022 r.

Modelowanie

Arkusz ocen

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