Komisja Egzaminacyjna dla Aktuariuszy

LXXXIV Egzamin dla Aktuariuszy

Sesja egzaminacyjna w dniu 12 kwietnia 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) 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.***) Scharakteryzuj efekt interakcji w uogólnionych modelach liniowych (GLM). Opisz na czym polega i jak jest uwzględniany.
b) (***1p.***) Czy można mówić o występowaniu efektu interakcji w modelach GLM, w których zmienne objaśniające nie są ze sobą skorelowane? Odpowiedź uzasadnij.
c) (***2p.***) Przedłużenie przez klienta zakładu ubezpieczeń polisy na kolejny rok modelowano za pomocą regresji logistycznej. Przyjęto, że zmienna zależna Y przyjmuje wartość: $Y = 1$, gdy klient przedłuży umowę oraz $Y = 0$, gdy nie przedłuży. Uwzględniono następujące zmienne objaśniające:

plec: Płeć (K - kobieta, M – mężczyzna)
lojalnosc: Liczba lat, w których kierowca był klientem zakładu.

Metodą największej wiarygodności oszacowano następujące parametry tego modelu:

	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-0.458	0.13364	0.00061
lojalnosc	0.045	0.01249	0.00027
*plec*M	-1.592	0.16972	0.00000
*lojalnosc*plec*M	0.046	0.01441	0.00133

Zdefiniuj co to jest szansa i oceń jak zmienia się szansa przedłużenia przez klienta polisy, gdy *lojalnosc* wzrasta o 1 rok.

Odpowiedzi:

.....

Odp. a)

Odpowiedź powinna zawierać co najmniej następujące informacje:

- Efekt interakcji pomiędzy zmiennymi objaśniającymi oznacza zmianę siły wpływu jednej ze zmiennych (na zmienną zależną) przy różnych wartościach innej zmiennej.
- Jest uwzględniany przez wprowadzenie do modelu sztucznego predyktora (sztucznej zmiennej objaśniającej) będącej iloczynem dwóch lub większej liczby zmiennych.

.....

Odp. b)

Tak, efekt interakcji może występować w modelach GLM, w których zmienne objaśniające nie są ze sobą skorelowane. Jako przykład można podać portfel ubezpieczeń OC o tej samej strukturze wiekowej dla kobiet i mężczyzn. Płeć i wiek są zmiennymi niezależnymi. Młodzi kierowcy płci męskiej są na ogół bardziej niebezpieczni w porównaniu z młodymi kierowcami płci żeńskiej. Prawidłowość ta znika lub odwraca się w starszym wieku. Tak więc, wiek i płeć mogą wchodzić w interakcje, mimo że są niezależne.

.....

Odp. c)
Szansą ($odds(A)$) nazywamy iloraz prawdopodobieństwa wystąpienia zdarzenia A do prawdopodobieństwa jego niewystąpienia:

$$odds(A) = \frac{P(A)}{1 - P(A)}$$

Szansa przedłużenia polisy:
- dla kobiet wzrost o 4.603%
- dla mężczyzn wzrost o 9.527%

Rozwiązanie:

$$\ln \frac{p_i}{1 - p_i} = \begin{cases} -0.458 + 0.045 \cdot lojalnosc, & \text{dla kobiet} \\ -0.458 - 1.592 + (0.045 + 0.046) \cdot lojalnosc, & \text{dla mężczyz} \end{cases}$$

Zadanie 2.

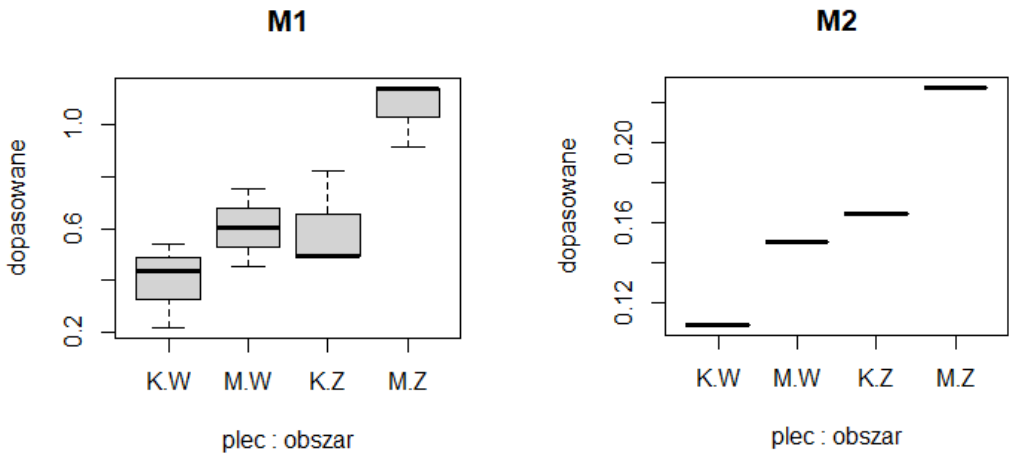
a) (***2p.***) Co to jest zmienna offsetowa (zmienna przesunięcia, offset) i jaką rolę odgrywa w szacowaniu uogólnionych modeli liniowych?

b) (***2p.***) W celu predykcji rocznej liczby szkód, na podstawie danych podanych w Tab. 2.1, oszacowano dwa modele regresji Poissona M1 i M2 z linkiem kanonicznym dla dwóch różnych zmiennych objaśnianych (zależnych). W obu modelach uwzględniono zmienną *E*, jednak występowała ona w różnych rolach. Uzyskano takie same wartości parametrów przy zmiennych objaśniających dla tych modeli. Na Rys. 2.1 przedstawiono w postaci wykresów pudełkowych wartości dopasowane (teoretyczne) dla M1 i M2.

Tabela 2.1

Nr polisy	Czas trwania (ekspozycja) w latach	Płeć (K-kobieta, M-mężczyzna)	Obszar zamieszkania	Liczba szkód
	E	*plec*	*obszar*	*K*
1	5	M	W	0
2	5	K	W	0
3	4	M	W	1
4	3	K	Z	1
5	4	K	W	0
6	3	K	Z	1
7	5	M	Z	0
8	5	M	Z	2
9	3	M	W	1
10	2	K	W	1
11	4	M	Z	1
12	5	K	Z	0

Rysunek 2.1



Scharakteryzuj model M1 i M2. W szczególności, podaj postacie predyktorów liniowych (prostych regresji) i wskaż jaką rolę w estymacji M1 i M2 odgrywa zmienna *E*.

c) (***1p.***)W tabeli Tab. 2.2 podano wartości oszacowanych parametrów. Wyznacz wartości dopasowane (teoretyczne) dla M1 i M2 dla grupy kierowców M.W (mężczyźni, obszar W). Jak je interpretujemy.

Tabela 2.2

	Estimate
(Intercept)	-2.2214
*plec*M	0.3282
*obszar*Z	0.4152

Odpowiedzi

.....

Odp. a)

Offset (przesunięcie) jest współzmienną uogólnionego modelu regresji o stałym parametrze równym 1, którego się nie szacuje. Reprezentuje zatem informacje a priori wnoszoną do modelu. Offset jest najczęściej używany do skalowania modelowania średniej w regresji Poissona z linkiem logarytmicznym.

Zmienna offsetowa uwzględnia ekspozycję na ryzyko. W przypadku modelowania liczby szkód K_i oznacza najczęściej czas trwania polisy. Uwzględnienie jej powoduje, że $E(K_i)$ zmienia się proporcjonalnie do ekspozycji. Innymi słowy, jeżeli czas trwania polisy rośnie, wówczas wartość oczekiwana $E(K_i)$ również rośnie.

.....

Odp. b)

Model:

M1 - predyktor liniowy: $\ln(K) = \beta_0 + \beta_1 \cdot plec + \beta_2 \cdot obszar + \ln(E)$, zmienna E występuje w roli offsetu.

M2 - predyktor liniowy: $\ln(\frac{K}{E}) = \beta_0 + \beta_1 \cdot plec + \beta_2 \cdot obszar$, zmienna E występuje w roli wag.

.....

Odp. c)

Model M1:

- polisa 1: $\widehat{K} = \exp(-2.2214 + 0.3282 + \ln(5)) = 0.7529458$
- polisa 3: $\widehat{K} = \exp(-2.2214 + 0.3282 + \ln(4)) = 0.6023566$
- polisa 9: $\widehat{K} = \exp(-2.2214 + 0.3282 + \ln(3)) = 0.4517675$

Wartości $\widehat{K}$ są to średnie dla liczby szkód w czasie ekspozycji odpowiednio 5, 4 i 3 lata.

Model M2:

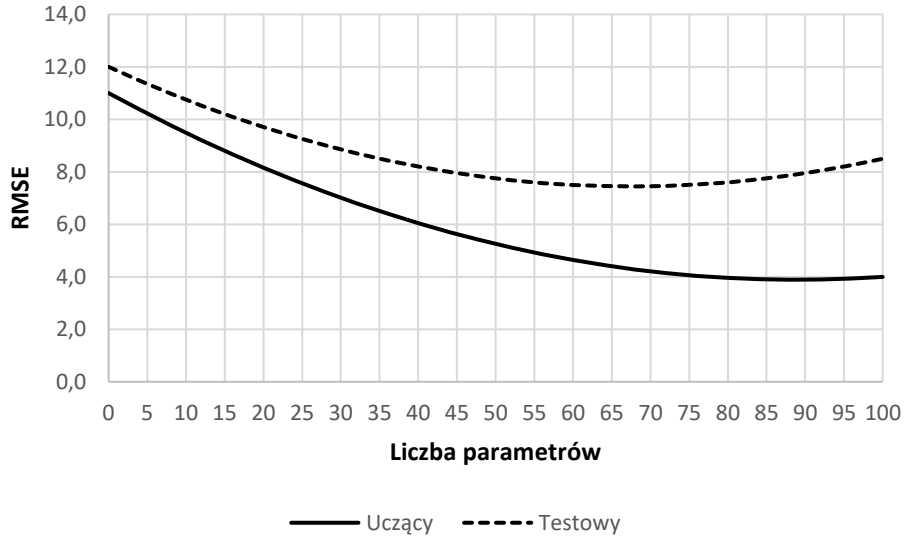
Dla wszystkich polis z grupy M.W (mężczyźni, obszar W): $\frac{\widehat{K}}{E} = \exp(-2.2214 + 0.3282) = 0.1505892$. Wartość $\frac{\widehat{K}}{E}$ to średnia roczna liczba szkód dla kierowców z grupy M.W.

Rozwiązanie:

Zadanie 3.

Aktuariusz podzielił dane na zbiór uczący i zbiór testowy. Następnie w oparciu zbiór uczący oszacował kilkadziesiąt modeli. Jakość tych modeli sprawdził na zbiorze uczącym i testowym, obliczając RMSE. Na poniższym wykresie (Rys.3.1) przedstawiono zależność między jakością modeli a ich złożonością reprezentowaną przez liczbę parametrów.

Rysunek 3.1



a) (***2p.***) Krótko opisz co najmniej dwa powody podziału danych na zbiór uczący i zbiór testowy.
b) (***2p.***) Krótko opisz złożoność i jakość następujących modeli:
   i. M1: 10 parametrów,
   ii. M2: 70 parametrów,
   iii. M3: 100 parametrów.

   Czy każdy z nich zachowuje optymalną równowagę między złożonością a wydajnością? Wskaż optymalny.
c) (***1p.***) Zidentyfikuj i krótko opisz jedną sytuację, w której korzystne jest dzielenie danych w szeregu czasowym na zbiory uczący i testowy według czasu, a nie w sposób losowy.

Odpowiedzi:

.....

Odp. a)

Na przykład:

1. Sprawdzenie modelu na danych, które nie były wykorzystywane do jego szacowania. Testowanie działania modelu na tym samym zestawie danych, na którym model został zbudowany, daje zbyt optymistyczne wyniki. Użycie danych uczących do porównania tego modelu z innymi modelami zbudowanymi na różnych danych daje mu nieuczciwą przewagę.

2. Sprawdzenie, czy model nie został przeuczony. Wraz ze wzrostem złożoności modelu dopasowanie do danych uczących będzie zawsze coraz lepsze. Natomiast w przypadku danych testowych, które nie brały udziału w procesie dopasowywania modelu, dodatkowa złożoność może nie poprawić jakości modelu. W miarę jak model staje się bardziej złożony, jego jakość na zbiorze testowym jest coraz słabsza, aż w końcu się pogorsza. Można to zaobserwować na rysunku w tym pytaniu.

.....

Odp. b)

Model M2 ma odpowiednią równowagę, ponieważ ma najmniejszy testowy RMSE. Model M1 jest zbyt prosty, natomiast model M3 jest zbyt złożony.

.....

Odp. c)

Na przykład zadbanie o to, aby w obydwu zbiorach, tj. uczącym i testowym można było wyodrębnić te same główne składowe szeregu czasowego (np. trend, wahania okresowe, wahania cykliczne,...).

Rozwiązanie:

Zadanie 4.
Modelując liczbę szkód w pewnym portfelu ubezpieczeń AC oszacowano dwa modele regresji Poissona: A i B. Przy czym w modelu A uwzględniono wszystkie zmienne z modelu B i dodatkowo zmienną jakościową *Marka.samoch* określającą markę samochodu. Wybrane wyniki oszacowań podano w Tab. 4.1.

Tabela 4.1

	Model A	Model B
Liczba oszacowanych parametrów	31	25
Dewiancja	62987.71	63059.85
Kryterium informacyjne AIC	81908.42	81968.57

a) (***1p.***) Wskaż liczbę kategorii zmiennej *Marka.samoch*.
b) (***2p.***) Co to jest dewiancja i dlaczego jest ważną miarą w modelowaniu statystycznym. Zapisz dewiancję w regresji Poissona.
c) (***2p***.) Sprawdź, czy marka samochodu jest istotna statystycznie w modelowaniu liczby szkód (w rozpatrywanym przypadku). Krótko scharakteryzuj zastosowany test i uzasadnij jego wybór.
Krótko omów zasadność wykorzystania marki samochodu jako czynnika ryzyka w modelowaniu liczby szkód w portfelach ubezpieczeń.

Odpowiedzi

.....

Odp. a)

Liczba kategorii zmiennej *Marka.samoch*: $31 - 25 + 1 = 7$

.....

Odp. b)

Dewiancja jest miarą tego, jak blisko wartości dopasowanych (teoretycznych) modelu są wartości obserwowane. Pożądany jest model z mniejszą dewiacją (chociaż ważne jest, aby uważać na nadmierne dopasowanie). Obserwacje, które mają duży udział w dewiancji, należy zbadać pod kątem błędów lub wartości odstających. Miara ta może być wykorzystany w procesie porównywania dwóch modeli zagnieżdżonych.
Dewiancja w regresji Poissona (z logarytmicznym linkiem):

$$\sum_{i=1}^{n} 2(y_i(\ln y_i - \ln \hat{\mu}_i) - (y_i - \hat{\mu}_i)),$$

gdzie $\hat{\mu}_i$ oznacza wartości dopasowane (teoretyczne).

.....

Odp. c)

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.
Hipoteza zerowa H0: Dwa modele są statystycznie takie same.
Statystyka testowa wyraża się wzorem:

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

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:

$$\chi^2 = D\left(y; \hat{\theta}^P\right) - D\left(y; \hat{\theta}^Q\right) = 63059.85 - 62987.71 = 72.14$$

Stopnie swobody: $31 - 25 = 6$
Wartość krytyczna np. na poziomie istotności 0.05 wynosi (z tablic): 12.592

Wniosek: Na poziomie istotności 0.05 hipotezę zerwą należy odrzucić, czyli marka samochodu jest istotna w modelowaniu liczby szkód.

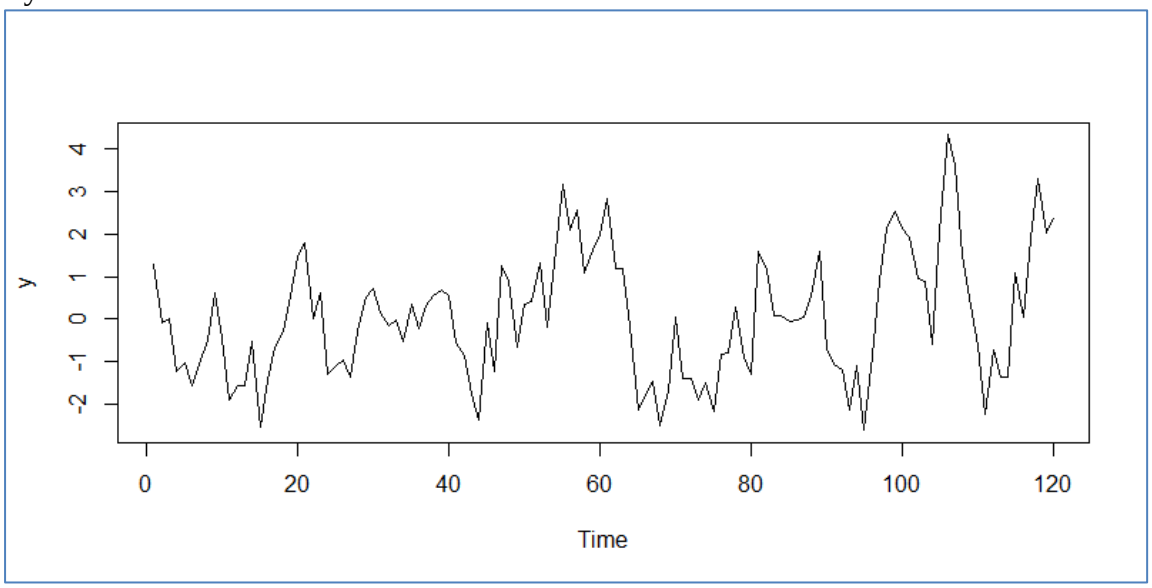
Z marką samochodu może być związany np. styl jazdy, różne marki to różne wartości samochodu (tym samym różna szkodowość), itp.

Rozwiązanie:

Zadanie 5.

Odchylenia miesięcznych kosztów całkowitych skorygowanych o inflację (w tys. zł) od wartości średniej w pewnej taryfie ubezpieczeniowej przedstawia stacjonarny szereg czasowy y_t (liczba obserwacji T=120) na rysunku 5.1. Z kolei, rysunek 5.2 prezentuje funkcje autokorelacji i autokorelacji cząstkowej dla tego szeregu.
Dla tego szeregu oszacowano dwa modele: M1- model ARIMA(1,0,0) i M2- model ARIMA(0,0,1). Wybrane wyniki są podane w tabeli 5.1.

Rysunek 5.1



Rysunek 5.2

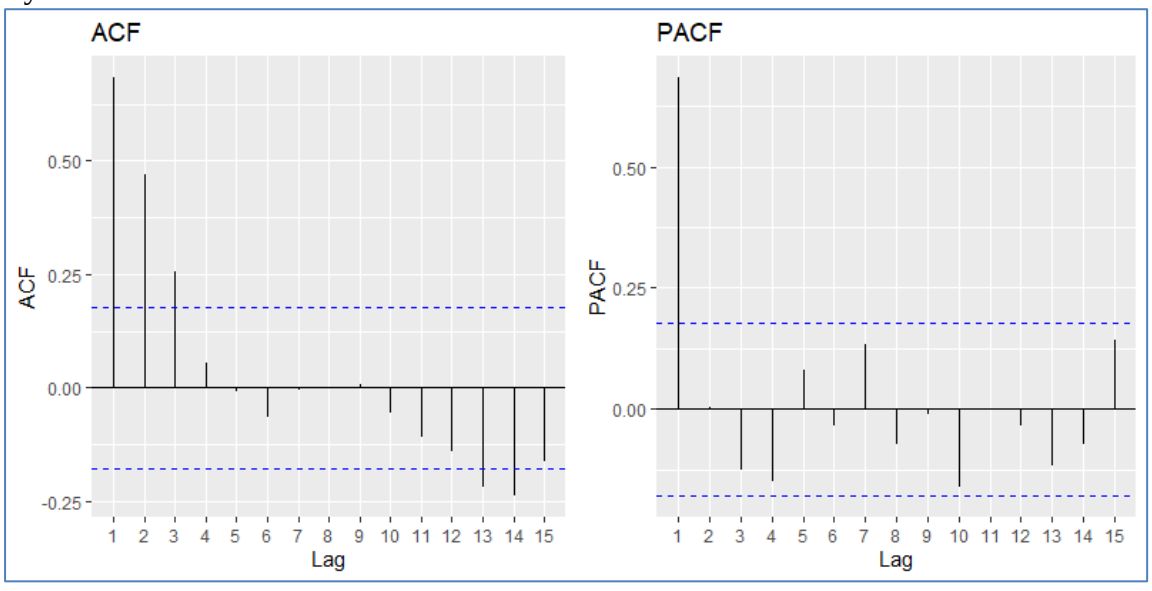


Tabela 5.1

<table>
<tr><th>Model M1 – ARIMA(1,0,0)</th><th>Model M2 – ARIMA(0,0,1)</th></tr>
<tr><td>Series: y
ARIMA(1,0,0) with zero mean
Coefficients:
ar1
0.6982
s.e. 0.0658

sigma^2 estimated as 1.111: log likelihood=-176.42
AIC=356.84 AICc=356.94 BIC=362.42

Ljung-Box test
data: Residuals from ARIMA(1,0,0) with zero mean
Q* = 9.5419, df = 9, p-value = 0.3888

Dodatkowe informacje o resztach tego modelu prezentuje Rys. 5.3

Wartości rzeczywiste i reszty dla 4 ostatnich miesięcy:
<table>
<tr><td>t</td><td>117</td><td>118</td><td>119</td><td>120</td></tr>
<tr><td>y_t</td><td>1.891</td><td>3.304</td><td>2.043</td><td>2.358</td></tr>
<tr><td>Reszty</td><td>1.860</td><td>1.984</td><td>-0.264</td><td>0.931</td></tr>
</table>
</td><td>Series: y
ARIMA(0,0,1) with zero mean
Coefficients:
ma1
0.6059
s.e. 0.0739

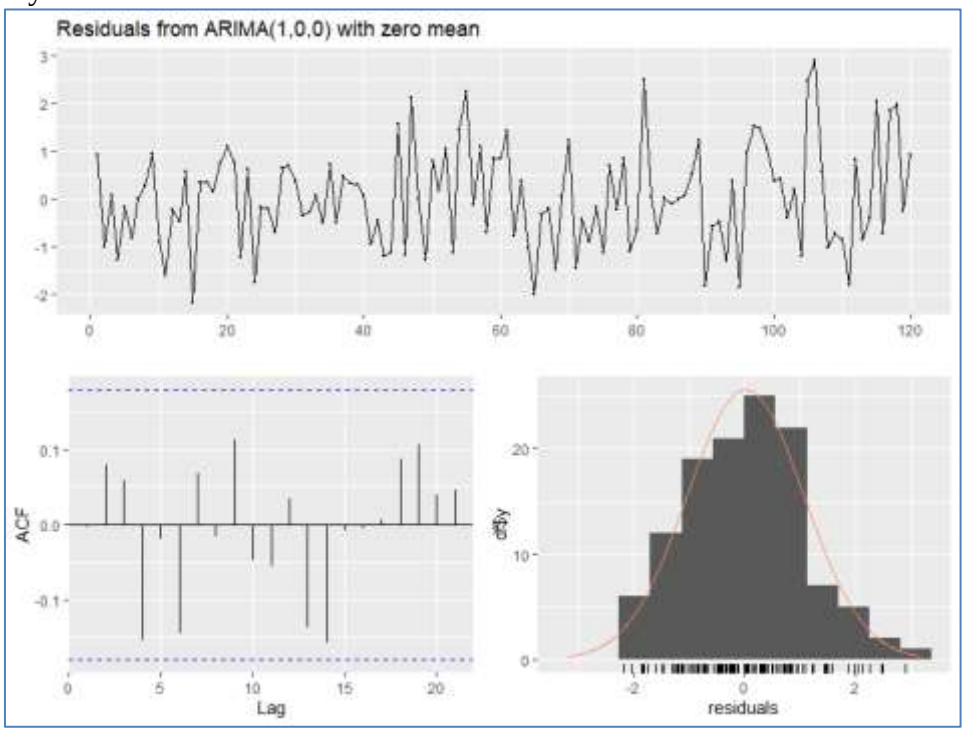
sigma^2 estimated as 1.367: log likelihood=-188.75
AIC=381.51 AICc=381.61 BIC=387.08

Ljung-Box test
data: Residuals from ARIMA(0,0,1) with zero mean
Q* = 32.528, df = 9, p-value = 0.0001613

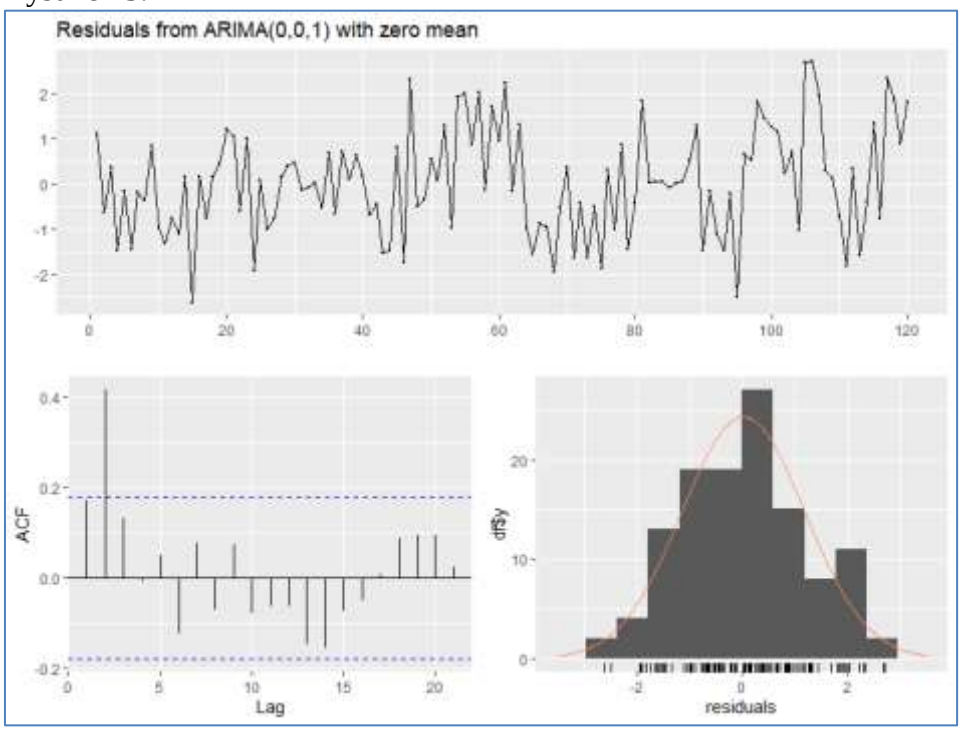
Dodatkowe informacje o resztach tego modelu prezentuje Rys. 5.4

Wartości rzeczywiste i reszty dla 4 ostatnich miesięcy:
<table>
<tr><td>t</td><td>117</td><td>118</td><td>119</td><td>120</td></tr>
<tr><td>y_t</td><td>1.891</td><td>3.304</td><td>2.043</td><td>2.358</td></tr>
<tr><td>Reszty</td><td>2.354</td><td>1.878</td><td>0.905</td><td>1.809</td></tr>
</table>
</td></tr>
</table>

Rysunek 5.3



Rysunek 5.4



a) (***2p.***) Krótko opisz funkcje autokorelacji i autokorelacji cząstkowej i ich przydatność we wstępnej identyfikacji modelu ARIMA.
b) (***2p.***) Wskaż, który z modeli M1 i M2 jest poprawny w analizowanym przypadku. Wybór uzasadnij, powołując się na Rys. 5.2 oraz wyniki z Tab. 5.1 (w tym na rysunki 5.3 i 5.4).
c) (***1p.***) Wykorzystując wskazany model wyznacz prognozy miesięcznego odchylenia y na okresy: $T+1,\ T+2$.

Odpowiedzi:

.....

Odp. a)

Autokorelacja i autokorelacja cząstkowa to miary związków między bieżącymi i przeszłymi wartościami szeregów określające, które przeszłe wartości szeregów są najbardziej użyteczne przy przewidywaniu przyszłych wartości.

Funkcja autokorelacji jest miarą korelacji między obserwacjami szeregu czasowego odległymi o k jednostek czasu (y_t i y_{t-k}).

Funkcja autokorelacji cząstkowej jest miarą korelacji między obserwacjami szeregu czasowego odległymi o k jednostek czasu (y_t i y_{t-k}), po skorygowaniu o obecność wszystkich pozostałych składników krótszego opóźnienia (y_{t-1}, y_{t-2}, ..., y_{t-k-1}).

Na podstawie wykresów tych funkcji można wstępnie określić rząd procesu autoregresji i średniej ruchomej.

.....

Odp. b)

Należało wskazać model M1 – ARIMA(1,0,0):

- funkcja autokorelacji znika (rys. 5.2),
- funkcja autokorelacji cząstkowej różni się istotnie od zera dla opóźnienia $k = 1$ (rys. 5.2),
- brak autokorelacji reszt modelu (tab. 5.1 - test Ljung-Boxa i wykres na rysunku 5.3),
- mniejsza wartość kryterium informacyjnego AIC (tab. 5.1).

.....

Odp. c)

Prognoza na $T + 1 = 121$:

$$y_{121}^{P} = 0.6982 \cdot 2.358 = 1.646356$$

Prognoza na $T + 1 = 122$:

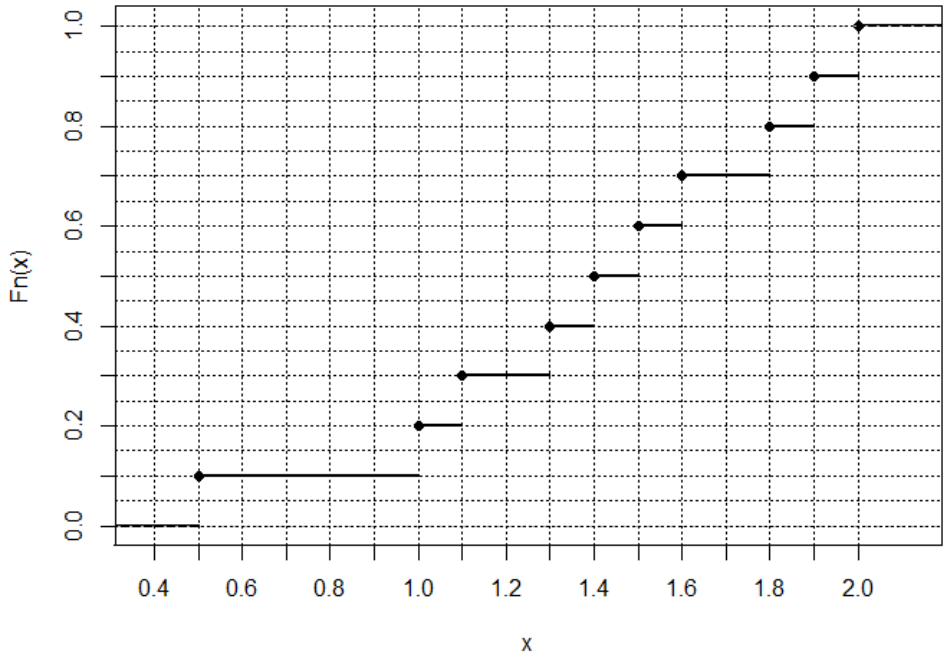
$$y_{121}^{P} = 0.6982 \cdot 1.646356 = \ 1.149486$$

Rozwiązanie:

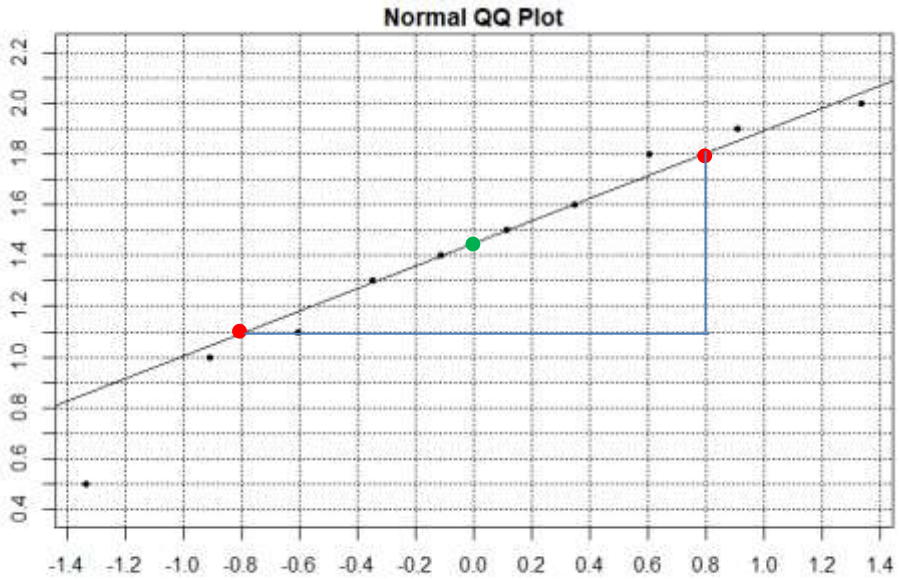
Zadanie 6.

Rozważamy wartości szkód z rozkładu lognormalnego. Rysunek 6.1 przedstawia dystrybuantę empiryczną dla logarytmów szkód. Rysunek 6.2 – wykres kwantylowy dla rozkładu normalnego dla logarytmów szkód. Wiadomo, że $\sum_{i=1}^{n} \ln(x_i) = 14.1$ oraz $\sum_{i=1}^{n} (\ln(x_i))^2 = 21.77$.

Rysunek 6.1



Rysunek 6.2



a) (***1p.***) Korzystając z dystrybuanty empirycznej wyznacz kwantyle rzędu: 0.90 i 0.95 rozkładu szkód.
b) (***1p.***) Wyjaśnij jakie wartości są na osiach przedstawionego na rysunku 6.2 wykresu kwantylowego i zinterpretuj ten wykres w kontekście rozważanego rozkładu lognormalnego.
c) (***2p.***) W przybliżeniu określ parametry narysowanej na rysunku 6.2 linii prostej. Jak interpretuje się te dwa parametry?
d) (***1p.***) Oszacuj wartość oczekiwaną wysokości szkód wykorzystując parametry oszacowane na podstawie wykresu kwantylowego oraz metodą największej wiarogodności.
   Wartość oczekiwana dla rozkładu lognormalnego: $E(X) = e^{\mu+0.5\cdot\sigma^2}$.

Odpowiedzi:

.....

Odp. a)

Dystrybuanta empiryczna:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{1\le i\le n} I(X_i \le x),$$

gdzie

$$I(X_i \le x) = \begin{cases} 1, \ gdy \ X_i \le x \\ 0, \ gdy \ X_i > x \end{cases}$$

Kwantyl rzędu α:

$$x_\alpha = \inf\{x \in R: \ \hat{F}_n(x) \ge \alpha \ \}$$

Kwantyle dla logarytmów szkód:

$$y_{0.90} = 1.9$$
$$y_{0.95} = 2.0$$

Stąd kwantyle dla szkód:

$$x_{0.90} = \exp(1.9) = 6.685894,$$
$$x_{0.95} = \exp(2.0) = 7.389056.$$

.....

Odp. b)

Na osi X są kwantyle standaryzowanego rozkładu normalnego $u_{\frac{i}{11}}$, gdzie $i = 1, \dots, 10$, natomiast na osi Y są uporządkowane wartości logarytmów szkód. Na podstawie wykresu można przyjąć, że logarytmy szkód mają rozkład normalny, czyli szkody podlegają rozkładowi lognormalnemu.

.....

Odp. c)

Wyraz wolny $a = 1.45$

Współczynnik kierunkowy b:

Przyjmując, że punkty o współrzędnych $(0.8, \ 1.8)$ oraz $(-0.8, \ 1.1)$ leżą na prostej, współczynnik kierunkowy jest równy:

$$b = \frac{1.8 - 1.1}{0.8 - (-0.8)} = 0.4375$$

Zatem prosta ma równanie:

$$y = 1.45 + 0.4375 \cdot x$$

Jeżeli przyjmiemy, że logarytmy szkód mają rozkład normalny, to wartość oczekiwaną i odchylenie standardowe (dla logarytmów szkód) można oszacować na poziomie odpowiednio 1.45 i 0.4375.

.....

Odp. d)

Na podstawie wykresu kwantylowego: $E(X) = e^{1.45+0.5\cdot 0.4375^2} = 4.691269$

Metodą największej wiarygodności:

- średnia dla logarytmów szkód: $\bar{y} = \frac{14.1}{10} = 1.41$
- odchylenie standardowe: $s = \sqrt{\frac{21.77}{9} - \frac{10\cdot 1.41^2}{9}} = 0.4581363$
- $E(X) = e^{1.41+0.5\cdot 0.4581363^2} = 4.549168$

Rozwiązanie:

Zadanie 7.

a) (***2p.***) Zmienna losowa X ma rozkład o dystrybuancie:

$$F(x) = \frac{e^x}{1+e^x}, x \in R.$$

Proszę podać algorytm symulacji wartości zmiennej X. Założyć, że znany jest sposób symulacji z rozkładu jednostajnego $U[0,1]$. Wyznaczyć realizację zmiennej X w przypadku wylosowania z $U[0,1]$ wartości $u = 0.4354$.

b) (***2p.***) Realizacje pewnej dwuwymiarowej kopuli C są symulowane w następujący sposób:
- Losujemy niezależnie dwie liczby u i v z rozkładu $U[0,1]$.
- Obliczmy:

$$w = \frac{u \cdot \sqrt{v}}{1-(1-u)\cdot\sqrt{v}}$$

- Tworzymy wektor (u, w).

Dany jest wektor losowy $\boldsymbol{X} = (X_1,\ X_2)$, dla którego:
- X_1 ma rozkład o dystrybuancie podanej w punkcie a),
- X_2 ma rozkład wykładniczy z parametrem $\lambda = 2$ (dystrybuanta rozkładu wykładniczego: $F(x) = 1 - e^{-\lambda x}$),
- kopula $C_{\boldsymbol{X}} = C$.

Proszę przestawiać algorytm symulacji wartości wektora $\boldsymbol{X}$. Wykorzystując ten algorytm, obliczyć realizację $\boldsymbol{X}$ przy założeniu, że ze zmiennej z $U[0,1]$ wylosowano niezależnie dwie wartości $u = 0.4354$ oraz $v = 0.8551$.

c) (***1p.***) Na podstawie dwóch niezależnie wylosowanych liczb ze standardowego rozkładu normalnego $z_1 = 0.5500$ i $z_2 = -0.3300$, proszę wyznaczyć realizację (symulację) wektora losowego o dwuwymiarowym standardowym rozkładzie normalnym ze współczynnikiem korelacji $\rho = 0.6$.

Odpowiedzi:

.....

Odp. a)

Algorytm:
1. Losujemy z rozkładu jednostajnego $U[0,1]$ liczbę u.
2. Wyznaczamy wartość $F^{-1}(u)$.

Realizacja:

$F^{-1}(u) = \ln\frac{u}{1-u}$.

$$F^{-1}(0.4354) = \ln\frac{0.4354}{1-0.4354} = -0.2598524$$

.....

Odp. b)

Algorytm:
1. Losujemy wektor (u, w) z kopuli C (w sposób opisany w punkcie b).
2. Wyznaczamy realizację (x_1, x_2) wektora $\boldsymbol{X} = (X_1,\ X_2)$ przyjmując:

$$x_1 = F_{X_1}^{-1}(u),$$
$$x_2 = F_{X_2}^{-1}(w).$$

Realizacja:

$$w = \frac{u \cdot \sqrt{v}}{1 - (1-u) \cdot \sqrt{v}} = \frac{0.4354 \cdot \sqrt{0.8551}}{1 - (1 - 0.4354) \cdot \sqrt{0.8551}} = 0.8424712$$

$$F_{X_1}^{-1}(u) = \ln \frac{0.4354}{1 - 0.4354} = -0.2598524$$

$$F_{X_2}^{-1}(w) = -\frac{\ln(1-w)}{\lambda} = -\frac{\ln(1 - 0.8424712)}{2} = 0.9240735$$

Czyli otrzymujemy: $(-0.2598524,\ 0.9240735)$

.....

Odp. c)

$$w_1 = z_1$$
$$w_2 = \rho \cdot z_1 + \sqrt{1-\rho^2} \cdot z_2$$

Stąd:

$$w_1 = 0.5500$$
$$w_2 = 0.6 \cdot 0.5500 + \sqrt{1 - 0.6^2} \cdot (-0.3300) = \ 0.066$$

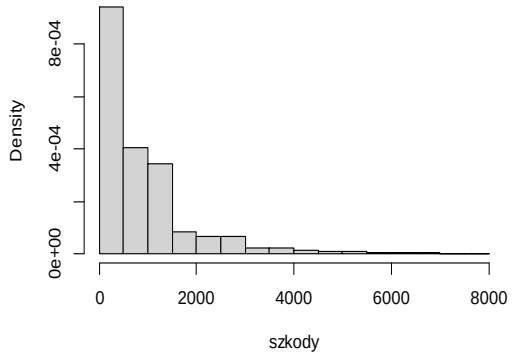
Czyli otrzymujemy wektor: $(0.5500,\ 0.0660)$

Rozwiązanie:

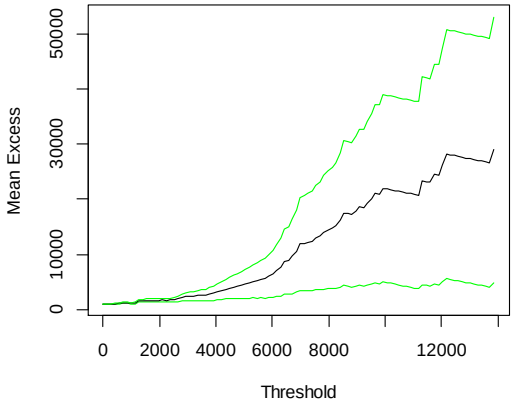
Zadanie 8.

a) (***2p***) Co to jest funkcja wartości oczekiwanej nadwyżki (*mean excess function*; inne spotykane nazwy tej funkcji: funkcja wartości oczekiwanej strat ponadprogowych, średnia funkcja nadwyżki, funkcja wartości oczekiwanej progu)? Podaj jej definicję, metodę estymacji i zastosowanie.
b) (*1**p***) Podaj przykład funkcji *mean excess function* dla wybranego jednego rozkładu.
c) (***2p***) Rysunki 8.1 i 8.2 przedstawiają odpowiednio histogram oraz empiryczną funkcję wartości oczekiwanej nadwyżki dla wysokości szkód w pewnym portfelu ubezpieczeń. Na tej podstawie krótko scharakteryzuj rozkład wysokości szkód. Wskaż co najmniej dwie własności tego rozkładu.

Rysunek 8.1



Rysunek 8.2



Odpowiedzi:

.....

Odp. a)

Niech Y będzie zmienną losową taką, że $E(Y) < \infty$. Funkcja wartości oczekiwanej nadwyżki (*mean excess function*) $e(\cdot)$ odpowiadająca Y jest zdefiniowana w następujący sposób:

$$e(u) = E(Y - u | Y > u).$$

Funkcja ta może być estymowana na podstawie próby losowej $\{y_1, y_2, \dots, y_n\}$ z wykorzystaniem następującego estymatora:

$$\hat{e}_n(u) = \frac{\sum_{i=1}^{n}(y_i - u) \cdot I(y_i > 0)}{\#\{y_i | y_i > u\}},$$

gdzie $\#\{y_i | y_i > u\}$ oznacza liczbę obserwacji przekraczających u, natomiast $I(\cdot)$ - funkcję indykatorową.

Funkcję $e(u)$ wykorzystuje się m.in. do określania zachowania rozkładu Y w ogonie, ustalania wartości progowej w analizie POT.

.....

Odp. b)

Na przykład dla zmiennej Y o rozkładzie wykładniczym z parametrem λ funkcja wartości oczekiwanej nadwyżki jest równa: $e(u) = E(Y) = \frac{1}{\lambda}$.

.....

Odp. c)

Rysunek 8.1 wskazuje na rozkład z silną asymetrią prawostronną, a rys. 8.2 na rozkład o ciężkim ogonie. Na podstawie rys. 8.2 można także wskazać wartość progową równą około 6000 w analizie POT.

Rozwiązanie:

Zadanie 9.

W poniższej tabeli (Tab. 9.1) podano zaobserwowane wartości dwuwymiarowej zmiennej losowej $(X,\ Y)$:

Tabela 9.1

Nr. obserwacji	*X*	*Y*
1	19	18
2	18	23
3	17	16
4	16	8
5	14	14
6	12	17

a) (***1p.***) Oblicz współczynnik korelacji liniowej Pearsona.
b) (***1p.***) Oblicz współczynnik korelacji rang Spearmana.
c) (***1p.***) Oblicz współczynnik korelacji rang Kendalla.
d) (***2p.***) Wymień wady współczynnika korelacji liniowej Pearsona.

Odpowiedzi:

.....

Odp. a)
Współczynnik korelacji liniowej Pearsona: 0.310535

.....

Odp. b)
Współczynnik korelacji rang Spearmana: 0.5428571

.....

Odp. c)
Współczynnik korelacji rang Kendalla: 0.3333333

.....

Odp. d)
Wady współczynnika korelacji liniowej Pearsona (np.):

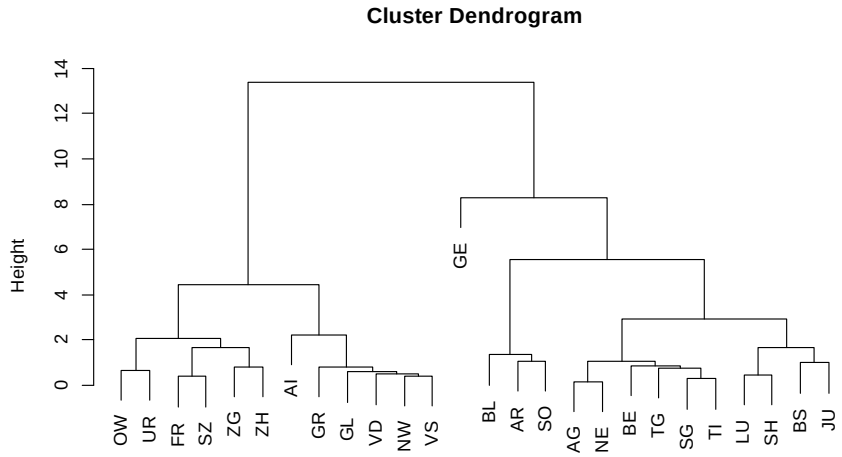
- Jest miarą zależności liniowej.
- Jest miarą zależności w wielowymiarowych rozkładach eliptycznych.
- Brak korelacji liniowej nie oznacza, że zmienne są niezależne.
- Zmienne losowe muszą mieć skończone wariancje, co jest szczególnie niepożądane dla niektórych rozkładów stosowanych w ubezpieczeniach, np. rozkładu Pareto.
- Wartość współczynnika korelacji nie jest niezmiennicza ze względu na ściśle monotoniczne przekształcenia nieliniowe.
- Przedział wartości przyjmowanych przez współczynnik korelacji zależy od rozkładów brzegowych.
- Korelacja nie oznacza związku przyczynowego.

Rozwiązanie:

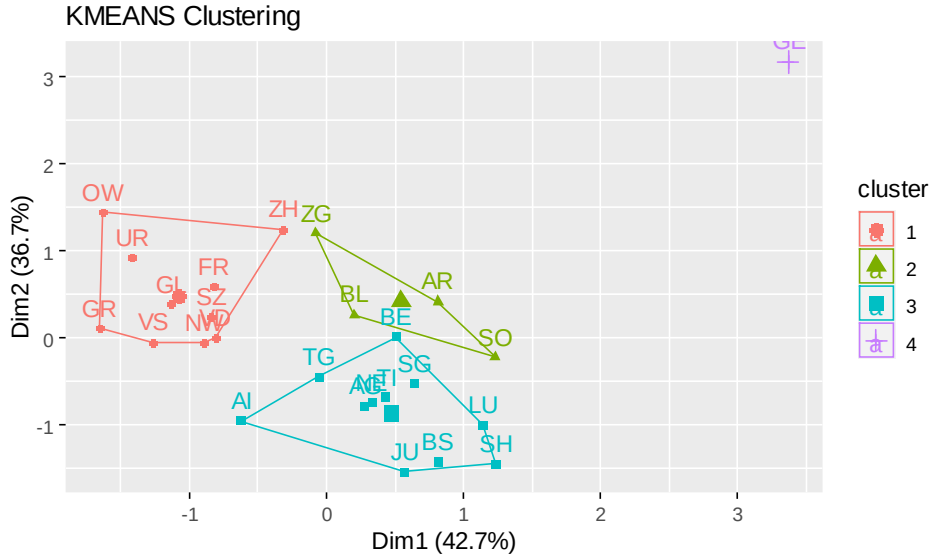
Zadanie 10.

Samochody kierowców tworzących portfel ubezpieczeń OC pewnego zakładu ubezpieczeń są zarejestrowane w 26 obwodach administracyjnych. Przeprowadzono klasyfikację (grupowanie) tych obwodów w oparciu 5 cech (zmiennych), które zdaniem aktuariusza mają wpływ na taryfę składek. Zastosowano dwie metody klasyfikacji: ***Warda*** i ***k-średnich*.** Wyniki grupowania przedstawiają odpowiednio Rys. 10.1 i Rys. 10.2. Z kolei na rysunkach 10.3 i 10.4 przedstawiono wskaźnik sylwetkowy (*silhouette index*) odpowiednio dla podziału na **trzy grupy** metodą Warda i na **cztery grupy** metodą k-średnich.

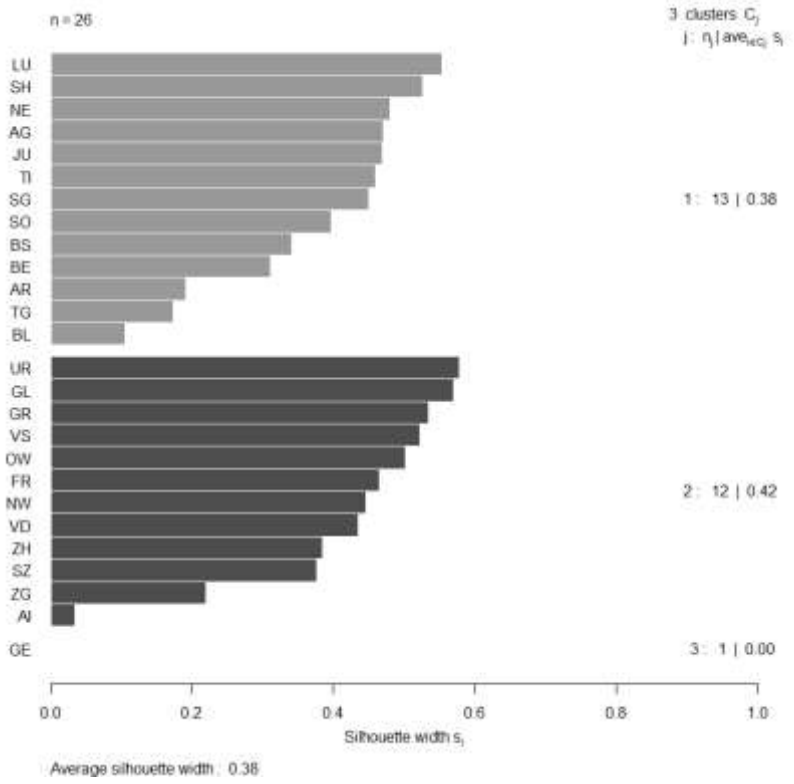
Rysunek 10.1



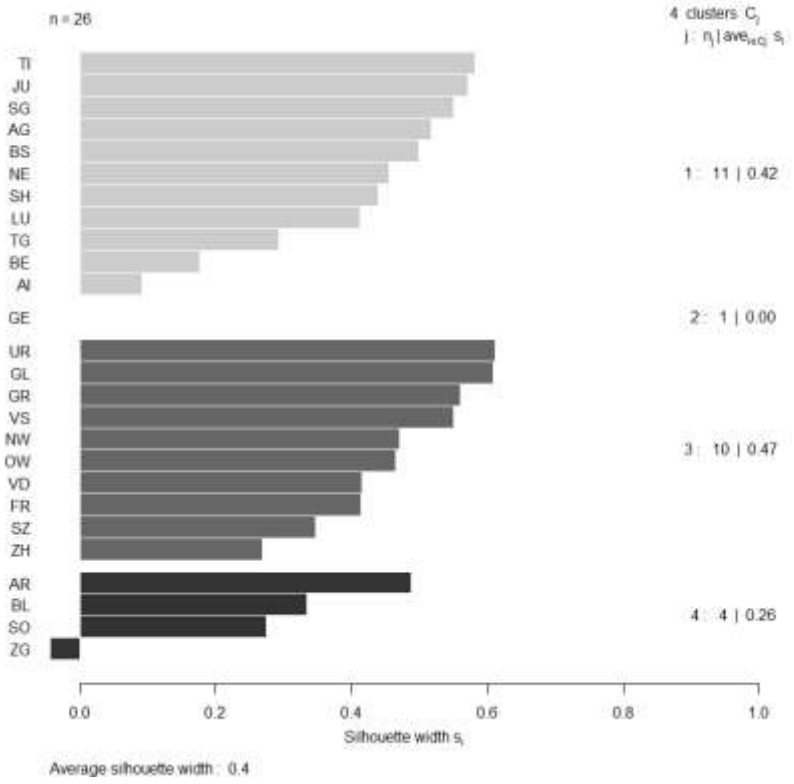
Rysunek 10.2



Rysunek 10.3. Metoda Warda



Rysunek 10.4 Metoda k-średnich



a) (***1p.***) Omów hierarchiczne i niehierarchiczne metody klasyfikacji (analizy skupień).
b) (***2p.***) Krótko opisz zastosowane w zadaniu metody klasyfikacji. Do jakich grup wskazanych w punkcie a) są zaliczane.
c) (***1p.***) Co to jest wskaźnik sylwetkowy (*silhouette index*) i w jakim celu jest wykorzystywany?
d) (***1p.***) Biorąc pod uwagę podane wyniki grupowania, wskaż, która z zastosowanych metod dała lepsze wyniki. Wybór uzasadnij.

Odpowiedzi:

.....

Odp. a)

Metody hierarchiczne. W procesie klasyfikacji wydziela się stopnie, na których badane obiekty łączone są w grupy (lub są rozdzielane). W wyniku analizy otrzymuje się hierarchiczną strukturę skupień. Jest ona najczęściej prezentowana w formie drzewa skupień (dendrogramu). Uzyskiwana hierarchia zezwala na dokładne określenie jak wzajemnie usytuowane są poszczególne skupienia oraz obiekty zawarte w wyodrębnionych skupieniach. W odróżnieniu od wyników uzyskiwanych na drodze stosowania metod niehierarchicznych, otrzymuje się strukturę skupień i strukturę obiektów uporządkowaną hierarchicznie, zgodnie z malejącym podobieństwem lub rosnącą odległością.

Metody niehierarchiczne . W technikach niehierarchicznych nie uwzględnia się porządku tworzenia grup. Obiekty, które znalazły się w jednej grupie, niekoniecznie muszą pozostawać razem. Obiekty mogą przechodzić z jednej grupy do innej. Metody te mają przewagę nad hierarchicznymi w analizie dużych zbiorów danych, dla których tworzenie dendrogramu byłoby procesem bardzo złożonym obliczeniowo.

.....

Odp. b)

Metoda Warda. Należy do grupy metod hierarchicznych. Jest to technika aglomeracyjna, w której początkowo każdy obiekt stanowi osobne skupienie, następnie obiekty leżące najbliżej siebie są łączone w nowe skupienie aż do uzyskania jednego skupienia. Do oszacowania odległości między skupieniami wykorzystuje podejście analizy wariancji. Zmierza do minimalizacji sumy kwadratów odchyleń wewnątrz skupień.

Metoda k-średnich. Należy do grupy metod niehierarchicznych. Grupowanie polega na podziale zbioru danych na z góry ustalone *k* skupień. Optymalna liczba skupień może zostać dopasowana po kilkukrotnym losowaniu próbki i tworzeniu na jej podstawie modelu. Kolejnym krokiem w analizie jest przenoszenie obiektów pomiędzy stworzonymi grupami tak, aby zminimalizować wariancję pomiędzy obiektami wewnątrz skupień oraz zmaksymalizować odległość pomiędzy skupieniami.

.....

Odp. c)
Wskaźnik sylwetkowy (*silhouette index*) służy do oceny rezultatów uzyskanej klasyfikacji w oparciu o informacje pochodzące z samego analizowanego zbioru. Jest definiowany dla każdego klasyfikowanego obiektu *i* w następujący sposób:

$$s_i = \frac{b(i) - a(i)}{\max(a(i),\ b(i))}$$

gdzie $a(i)$ to średnia odległość pomiędzy obiektem *i* a pozostałymi w tym samym skupieniu, natomiast $b(i)$ to średnia odległość obiektu *i* od najbliższego skupienia (do którego *i* nie należy). Miara ta przyjmuje wartości od -1 do 1. Obiekty, dla których s_i jest bliskie 1 są poprawnie sklasyfikowane, jeśli s_i jest bliskie 0 obiekt leży na granicy skupień, jeśli natomiast s_i jest ujemne obiekt znajduje się w złym skupieniu.
Średnia wartość tego wskaźnik dla obiektów z danego skupienia k ($\bar{s}_k$), wskazuje jak dobrze obiekty są przydzielone do tego skupienia. Natomiast średnia wartość dla wszystkich klasyfikowanych obiektów służy jako miara jakości otrzymanej struktury grupowej.

.....

Odp. d)
Można było wskazać metodę Warda lub k-średnich. Za pierwszą przemawia homogeniczność skupisk (grup) a dokładniej brak źle sklasyfikowanych obiektów, za drugą, wyższa wartość wskaźnika sylwetkowego.

Rozwiązanie:

Sesja egzaminacyjna w dniu 12 kwietnia 2022 r.

Modelowanie

Arkusz ocen

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