

**Komisja Egzaminacyjna dla Aktuariuszy**

**LXXXVII Egzamin dla Aktuariuszy**

**Sesja egzaminacyjna w dniu 24 stycznia 2023 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  $\chi^2_{\alpha;v}$  rozkładu chi-kwadrat spełniające warunek  $P(\chi^2 \geq \chi^2_{\alpha;v}) = \alpha$

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

**Zadanie 1.**

Wysokość pojedynczego roszczenia  $Y_i$  w pewnym portfelu ubezpieczeń AC modelowano z uwzględnieniem dwóch zmiennych objaśniających:

- *CarAge.kat* - wiek samochodu. Zmienna jakościowa przyjmująca następujące kategorie: [0, 5], (5, 10] i (10,100].
- *Gas* – rodzaj silnika. Zmienna jakościowa przyjmująca następujące kategorie: *Regular* i *Diesel*.

Oszacowano dwa uogólnione modele liniowe, w których uwzględniono powyższe zmienne objaśniające i założono rozkład gamma dla  $Y_i$ . Model M1, w którym nie uwzględniono interakcji między zmiennymi *CarAge.kat* i *Gas* oraz M2, w którym uwzględniono interakcje. W obydwu modelach przyjęto kanoniczną funkcję łączącą. Uzyskano następujące wyniki:

- **Model M1** (bez interakcji)

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	7.445e-04	9.683e-06	76.893	< 2e-16 ***
CarAge.kat(5,10]	2.867e-05	1.314e-05	2.183	0.029085 *
CarAge.kat(10,100]	7.396e-05	1.387e-05	5.332	9.83e-08 ***
GasRegular	4.219e-05	1.127e-05	3.745	0.000181 ***

---

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

(Dispersion parameter for Gamma family taken to be 0.7834687)

Null deviance: 11904 on 15866 degrees of freedom

Residual deviance: 11867 on 15863 degrees of freedom

AIC: 257024

- **Model M2** (z interakcją)

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	7.590e-04	1.127e-05	67.337	< 2e-16 ***
CarAge.kat(5,10]	2.348e-05	1.744e-05	1.347	0.178
CarAge.kat(10,100]	1.988e-05	1.869e-05	1.063	0.288
GasRegular	7.362e-06	1.725e-05	0.427	0.669
CarAge.kat(5,10]:GasRegular	1.327e-05	2.648e-05	0.501	0.616
CarAge.kat(10,100]:GasRegular	1.143e-04	2.768e-05	4.131	3.64e-05 ***

---

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

(Dispersion parameter for Gamma family taken to be **0.781385**)

Null deviance: 11904 on 15866 degrees of freedom

Residual deviance: 11852 on 15861 degrees of freedom

AIC: 257007

## – Analiza dewiencji

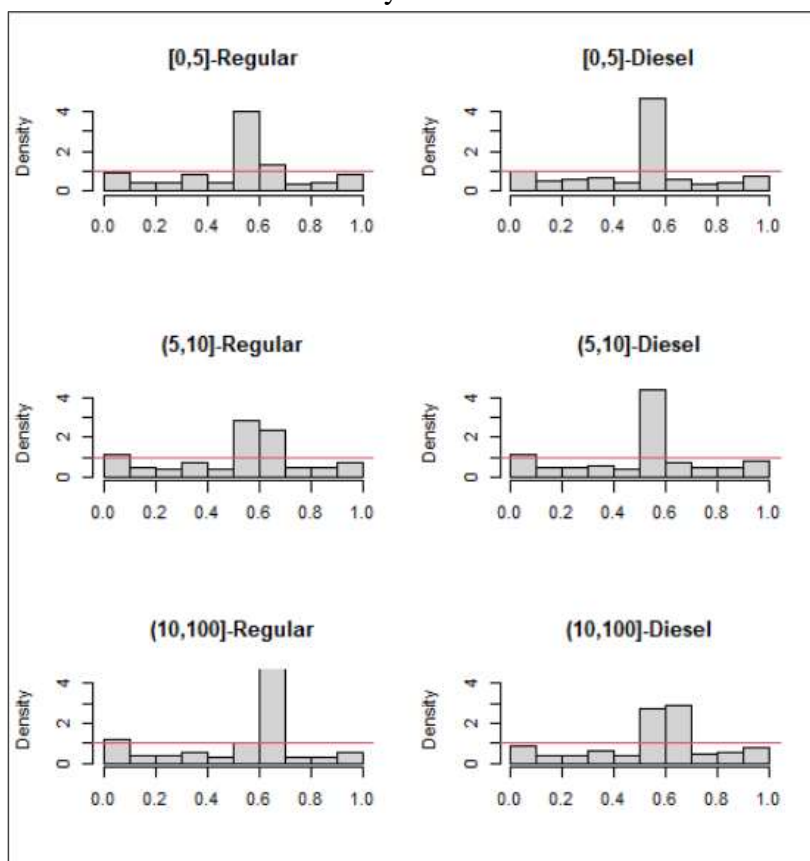
	Df	Deviance	Resid. Df	Resid. Dev	F	Pr(>F)
NULL			15866	11904		
CarAge.kat	2	25.886	15864	11878	16.564	6.511e-08 ***
Gas	1	11.024	15863	11867	14.108	0.0001732 ***
CarAge.kat:Gas	2	<b>14.606</b>	15861	11852	?	8.780e-05 ***

---

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

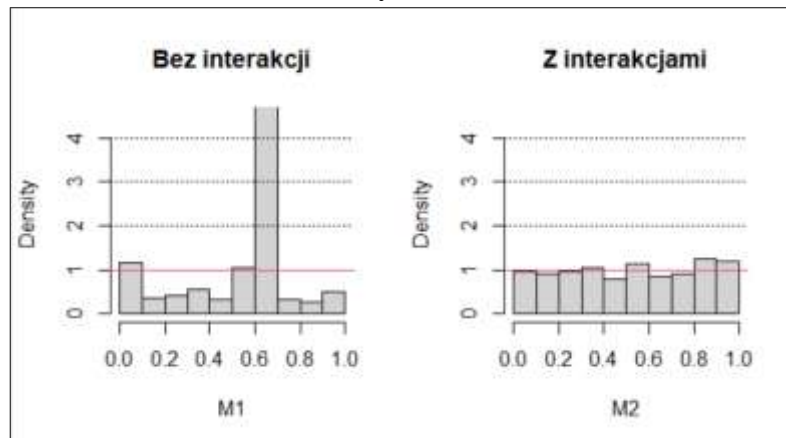
Wykorzystując zbiór testowy i model **M1**, dla każdej klasy ryzyka (tj. każdej kombinacji kategorii zmiennej *CarAge.kat* i *Gas*) skonstruowano histogram wartości  $F_{G_i}(y_{j,G_i})$ , gdzie:  $F_{G_i}$  – oszacowana dystrybuanta rozkładu gamma odpowiadająca klasie  $G_i$ ,  $y_{j,G_i}$  – wysokości zanotowanych roszczeń w klasie  $G_i$ . Histogramy są przedstawione na rysunku 1.1.

Rys.1.1.



Na rysunku 1.2 przedstawiono analogiczne wykresy dla klasy (10,100]-Regular skonstruowane z wykorzystaniem modelu M1 (z lewej strony) i Modelu M2 (z prawej strony).

Rys.1.2



- (1p.) Czy skonstruowane wykresy są przydatne w ocenie jakości oszacowanych modeli? Odpowiedź uzasadnij!
- (1p.) Uzasadnij dlaczego w analizie dewiancji wybrano test F. Oblicz brakującą wartość statystyki F dla interakcji.
- (2p.) Na podstawie podanych wyników wybierz lepszy model. Wybór uzasadnij odwołując się do wyników analizy dewiancji i podanych wykresów.
- (1p.) Wykorzystując wybrany model, wyznacz prognozę wysokości pojedynczego rozszczenia dla klasy: (10,100]-Regular.

### Odpowiedzi

#### Odp. a)

Tak.

W uzasadnieniu należało odwołać się do 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  $(0,1)$ .

#### Odp. b)

W uzasadnieniu należało wskazać, że w testowanym modelu był szacowany parametr dyspersji.

Statystyka wyraża się wzorem:

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

gdzie:

$D(y; \hat{\theta}^p)$  – dewiancja modelu o mniejszej liczbie parametrów  $p$ ,

$D(y; \hat{\theta}^q)$  – dewiancja modelu o większej liczbie parametrów  $q$ ,

---

$\hat{\phi}$  - oszacowanie parametru dyspersji modelu o większej liczbie parametrów.

Stąd (podstawiane wartości zaznaczono w treści zadania na czerwono):

$$F = \frac{14.606}{0.781385 \cdot 2} = 9.346225.$$

.....  
**Odp. c)**

Należało wskazać model M2. Wynik testu  $F$  wskazuje, że ma on lepsze zdolności predykcyjne, co potwierdza także wykres na rysunku 1.2.

.....  
**Odp. d)**

Prognoza

$$y^P = \frac{1}{0.000759 + 0.00001988 + 0.000007362 + 0.0001143} = 110.44$$

---

**Rozwiązanie:**

**Zadanie 2.**

Ubezpieczyciel chce zaoferować swoim dotychczasowym klientom nowe ubezpieczenie podróży z ochroną Covid. Chcąc dowiedzieć się, jaki wpływ miały określone cechy klienta na zawarcie ubezpieczenia podróży, aktuariusz na podstawie istniejących danych historycznych oszacował dwa modele logistyczne: model A i model B. Zmienna zależna w tych modelach przyjmuje dwie wartości:  $Y=1$ , gdy klient kupi ubezpieczenie oraz  $Y=0$ , gdy nie kupi). Aktuariusz wziął pod uwagę następujące cechy:

- *plec*: Płeć klienta (K - kobieta, M – mężczyzna)
- *zamieszkanie*: Zamieszkanie klienta (Miasto, Wies)
- *wiek*: Wiek klienta (zmienna jakościowa przyjmująca trzy kategorie: G1, G2, G3, przy czym im wyższa kategoria, tym klienci są starsi).
- *dochod*: Roczny dochód klienta w tys. zł (zmienna ilościowa).

Uzyskał następujące oszacowania:

– **Model A**

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-4.272412	0.300839	-14.202	< 2e-16 ***
plecM	-1.182367	0.129969	-9.097	< 2e-16 ***
zamieszkanieWies	-0.776746	0.115418	-6.730	1.7e-11 ***
wiekG2	0.375125	0.278850	1.345	0.179
wiekG3	1.508494	0.158664	9.507	< 2e-16 ***
dochod	0.043211	0.002582	16.734	< 2e-16 ***

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2402.7 on 1864 degrees of freedom

Residual deviance: 1894.8 on 1859 degrees of freedom

AIC: 1906.8

– **Model B**

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-3.627971	0.322008	-11.267	< 2e-16 ***
plecM	-1.166195	0.131567	-8.864	< 2e-16 ***
zamieszkanieWies	-0.788570	0.116191	-6.787	1.15e-11 ***
wiekG2	-1.053609	1.496586	-0.704	0.4814
wiekG3	-1.741291	0.755040	-2.306	0.0211 *
dochod	0.037605	0.002782	13.519	< 2e-16 ***
wiekG2:dochod	0.012794	0.013458	0.951	0.3418
wiekG3:dochod	0.035700	0.008208	4.350	1.36e-05 ***

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2402.7 on 1864 degrees of freedom

Residual deviance: 1871.6 on 1857 degrees of freedom

AIC: 1887.6

- a) (1p.) Zinterpretuj parametry **modelu A** dla zmiennej *wiek* i *dochod*. Wykorzystaj ilorazy szans.
- b) (1p.) Oblicz maksymalne prawdopodobieństwo kupna ubezpieczenia dla klienta z rocznym dochodem w wysokości 100 000 zł, wykorzystując model A i model B.
- c) (1p.) Roczny dochód klienta aktuariusz uwzględnił w modelu w sposób liniowy jako zmienną ilościową. Wymień trzy inne opcje.
- d) (2p.) Wykorzystując odpowiedni test, sprawdź czy wszystkie parametry (łącznie) interakcji między wiekiem a rocznym dochodem są statystycznie istotne. Zapisz odpowiednie hipotezy (tj. zerową i alternatywną), podaj statystykę testową i jej rozkład. Przyjmij poziom istotności równy 0.05.

### Odpowiedzi

#### Odp. a)

Zmienna *wiek*:

Parametr stojący przy kategorii *wiekG2* wynosi 0.375125. Dodatnia wartość parametru wskazuje, że klienci z kategorii wiekowej G2 są bardziej podatni na zawarcie ubezpieczenia podróży w porównaniu z klientami z kategorii wiekowej G1. Jeżeli klienci różnią się tylko kategorią wiekową, to klienci z kategorii G2 mają o  $(\exp(0.375125) - 1) \cdot 100\% = 45.52\%$  wyższe szanse na zawarcie ubezpieczenia podróży w porównaniu z klientami z kategorii wiekowej G1.

Parametr stojący przy kategorii *wiekG3* wynosi 1.508494. Dodatnia wartość parametru wskazuje, że klienci z kategorii wiekowej G3 są bardziej podatni na zawarcie ubezpieczenia podróży w porównaniu z klientami z kategorii wiekowej G1. Jeżeli klienci różnią się tylko kategorią wiekową, to klienci z kategorii G3 mają o  $(\exp(1.508494) - 1) \cdot 100\% = 351.99\%$  wyższe szanse na zawarcie ubezpieczenia podróży w porównaniu z klientami z kategorii wiekowej G1.

Zmienna *dochod*:

Parametr stojący przy zmiennej *dochod* wynosi 0.043211. Spośród klientów identycznych pod względem innych cech zawartych w modelu, a różniący się tylko rocznym dochodem o 1 tys. zł, klienci bogatsi (o 1 tys. zł.) mają  $(\exp(0.043211) - 1) \cdot 100\% = 4.42\%$  wyższe szanse na zawarcie ubezpieczenia podróży w porównaniu z klientami biedniejszymi.

#### Odp. b)

Maksymalne prawdopodobieństwo występuje w grupie kobiet mieszkających w mieście z kategorii wiekowej G3.

Model A:

$$\hat{p} = \frac{\exp(-4.272412 + 1.508494 + 0.043211 \cdot 100)}{1 + \exp(-4.272412 + 1.508494 + 0.043211 \cdot 100)} = 0.825952$$

Model B:

$$\hat{p} = \frac{\exp(-3.627971 - 1.741291 + 0.037605 \cdot 100 + 0.0357 \cdot 100)}{1 + \exp(-3.627971 - 1.741291 + 0.037605 \cdot 100 + 0.0357 \cdot 100)} = 0.87666$$



---

.....

**Odp. c)**

- założyć zależność wielomianową, np. kwadratową:  $dochod + dochod^2$
  - funkcja gładka (model GAM)
  - przekształcić na zmienną jakościową (utworzyć kategorie dochodu)
- .....

**Odp. d)**

Współczynniki interakcji (łącznie) są statystycznie istotne.

$H_0$ : Wszystkie współczynniki interakcji między wiekiem a rocznym dochodem są równe zero.

$H_1$ : Przynajmniej jeden współczynnik jest różny od zera.

Statystyka 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  $q$ ,

Stąd (podstawiane wartości zaznaczono w treści zadania na czerwono):

$$\chi^2 = 1894.8 - 1871.6 = 23.2$$

*Wartość statystyki można było także wyznaczyć korzystając wartości kryterium AIC dla tych modeli.*

Wartość krytyczna:  $\chi^2_{0.05;2} = 5.991$ .

---

**Rozwiązanie:**

**Zadanie 3.**

Ubezpieczyciel chce wykorzystać model regresji Poissona do analizy liczby szkód w swoim portfelu ubezpieczeń komunikacyjnych. Zebrał następujące dane dotyczące liczby roszczeń  $k_i$ ,  $i = 1, 2, \dots, 35$  z trzech różnych klas polis:

– Klasa A ( $i = 1, \dots, 10$ ): 1 2 0 2 1 0 0 2 2 1

– Klasa B ( $i = 11, \dots, 15$ ): 1 0 1 1 0

– Klasa C ( $i = 16, \dots, 35$ ): 0 0 0 0 0 1 0 1 0 0  
1 0 1 0 0 0 0 0 0 0

a) (2p.) Wykaż, że rozkład Poissona należy do wykładniczej rodziny rozkładów.

Przyjmij następującą parametryzację wykładniczej rodziny rozkładów:

$$f(y_i; \theta_i, \phi) = \exp\left(\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right),$$

gdzie:  $\theta_i$  - parametr kanoniczny,  $\phi$  - parametr dyspersji.

b) (3p.) Aktuariusz wybrał model, w którym:

$$\ln(\lambda_i) = \begin{cases} \alpha, & i = 1, \dots, 10 \\ \beta, & i = 11, \dots, 15 \\ \gamma, & i = 16, \dots, 35 \end{cases}$$

gdzie  $\lambda_i$  jest wartością oczekiwaną odpowiedniego rozkładu Poissona. Wyznacz logarytm funkcji wiarygodności dla tego modelu, a następnie znajdź oszacowania największej wiarygodności dla  $\alpha, \beta$  i  $\gamma$ .

**Odpowiedzi****Odp. a)**

Funkcja prawdopodobieństwa dla rozkładu Poissona ma postać:

$$f(y; \lambda) = \frac{\lambda^y e^{-\lambda}}{y!} = \exp(y \ln(\lambda) - \lambda - \ln(y!)),$$

stąd:

$$\theta = \ln(\lambda),$$

$$\phi = 1, \text{ czyli } a(\phi) = 1,$$

$$b(\theta) = e^\theta,$$

$$c(y, \phi) = -\ln(y!).$$

**Odp. b)**

Logarytm wiarygodności można zapisać w następujący sposób:

$$l(\lambda_A, \lambda_B, \lambda_C) = \ln(L(\lambda_A, \lambda_B, \lambda_C)) = \sum y_i \ln(\lambda_i) - \sum \lambda_i - \sum \ln(y_i!).$$

Stąd otrzymujemy:

---

$$\begin{aligned}l &= \alpha \sum_{i=1}^{10} y_i + \beta \sum_{i=11}^{15} y_i + \gamma \sum_{i=16}^{35} y_i - 10e^\alpha - 5e^\beta - 20e^\gamma - \sum_{i=1}^{35} \ln(y_i!) \\ &= 11\alpha + 3\beta + 4\gamma - 10e^\alpha - 5e^\beta - 20e^\gamma - \sum_{i=1}^{35} \ln(y_i!)\end{aligned}$$

Obliczając pochodne cząstkowe względem  $\alpha, \beta$  oraz  $\gamma$  otrzymujemy:

$$\frac{\partial}{\partial \alpha} l = 11 - 10e^\alpha$$

$$\frac{\partial}{\partial \beta} l = 3 - 5e^\beta$$

$$\frac{\partial}{\partial \gamma} l = 4 - 20e^\gamma$$

stąd:

$$\hat{\alpha} = \ln 1.1 = 0.09531$$

$$\hat{\beta} = \ln 0.6 = -0.51083$$

$$\hat{\gamma} = \ln 0.2 = -1.60944$$

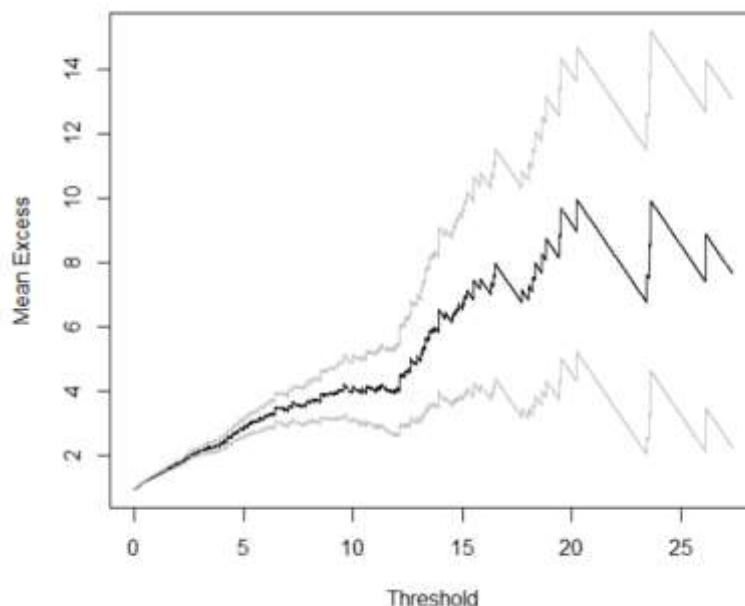
---

**Rozwiązanie:**

**Zadanie 4.**

- a) (1p.) Krótko przedstaw czym zajmuje się Teoria Wartości Ekstremalnych (*Extreme Value Theory, EVT*) i wskaż możliwości jej wykorzystania przez aktuarium.
- b) (1p.) Jednym z głównych podejść wykorzystywanych w *EVT* jest analiza przekroczeń progu *POT (Peaks Over Threshold)*. Krótko omów to podejście.
- c) (2p.) Wskaż dlaczego w metodzie *POT* kluczową rolę odgrywa ustalenie progu przekroczeń na właściwym poziomie. Uzasadnij dlaczego w tym celu (ustaleniu progu) można wykorzystać funkcję wartości oczekiwanej nadwyżki (*mean excess function*).
- d) (1p.) Na poniższym rysunku (Rys. 4.1.) przedstawiono empiryczną funkcję wartości oczekiwanej nadwyżki dla pewnego zbioru szkód. Na jej podstawie ustal wartość progu. Wybór uzasadnij!

Rys. 4.1.

**Odpowiedzi****Odp. a)**

Zobacz np.:

- Podrozdział 9.1 w: “Effective Statistical Learning Methods for Actuaries I” - M. Denuit, D. Hainaut, J. Trufin, Springer, 2019.
- Wstęp do rozdziału 5 (str. 135) w: “Quantitative Risk Management: Concepts, Techniques and Tools”, revised edition - A. McNeil, R. Frey, P. Embrecht, Princeton, 2015

---

.....

**Odp. b)**

Zobacz np. podrozdział 9.5 w: “Effective Statistical Learning Methods for Actuaries I” - M. Denuit, D. Hainaut, J. Trufin, Springer, 2019.

.....

**Odp. c)**

W odpowiedzi należało wskazać:

- na klasyczny problem związany z kompromisem obciążenie-wariancja: wybór zbyt niskiego progu oznacza, że założenie dotyczące ogona jest niewłaściwe, natomiast wybór zbyt wysokiego progu oznacza, że mamy zbyt mało obserwacji, aby rozsądnie oszacować parametry rozkładu. Szczegóły np. w podrozdziale 9.5.5 w: “Effective Statistical Learning Methods for Actuaries I” - M. Denuit, D. Hainaut, J. Trufin, Springer, 2019.
  - na liniową postać funkcji wartości oczekiwanej nadwyżki po przekroczeniu odpowiedniego progu. Szczegóły np. w podrozdziale 5.2.2 w: “Quantitative Risk Management: Concepts, Techniques and Tools”, revised edition - A. McNeil, R. Frey, P. Embrecht, Princeton, 2015.
- .....

**Odp. d)**

Należało wskazać wartość ok. 12. W uzasadnieniu powołać się na w przybliżeniu liniową postać empirycznej funkcji wartości oczekiwanej nadwyżki po przekroczeniu 12.

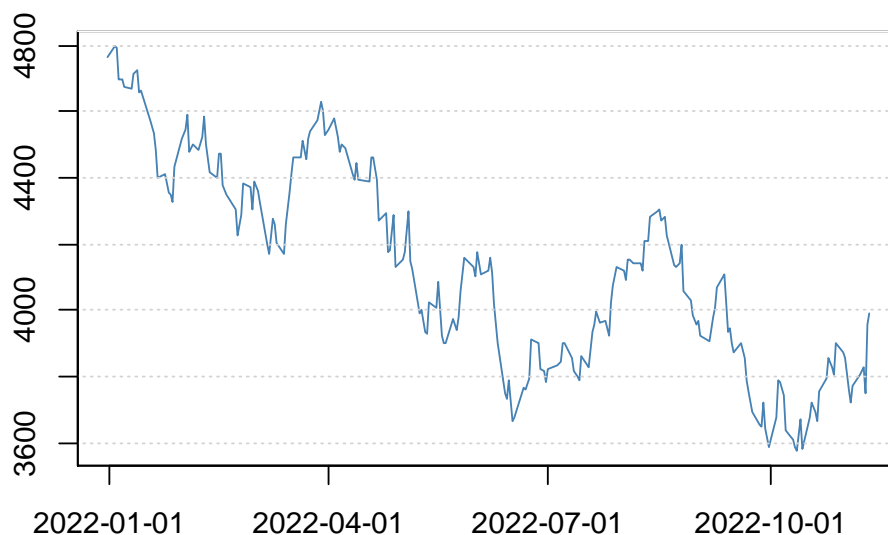
---

**Rozwiązanie:**

**Zadanie 5.**

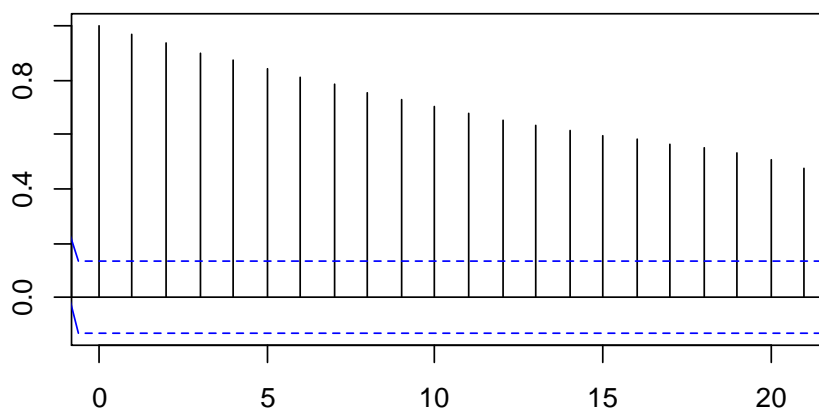
- (1p.) Wymień warunki stacjonarności (słabej) szeregu czasowego.
- (1p.) Podaj definicję procesu błędzenia losowego.
- (1p.) Wskaż i krótko przedstaw co najmniej dwie metody identyfikacji procesu błędzenia losowego.
- (2p.) Na poniższym wykresie (Rys. 5.1) podano notowania indeksu SP500 w okresie od 31-12-2021 do 11-11-2022.

Rys. 5.1. Notowania SP500



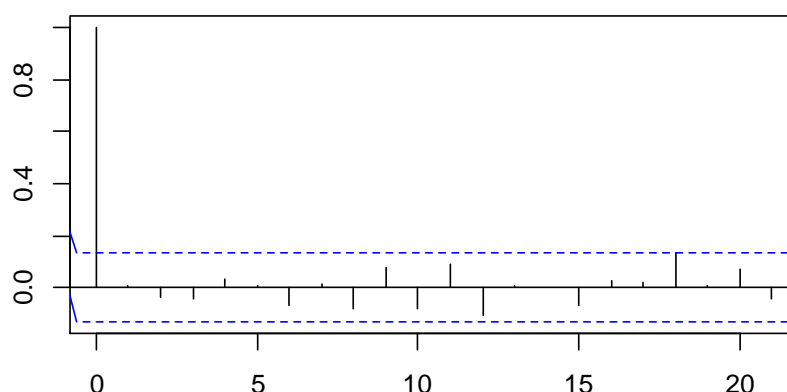
Funkcję autokorelacji dla tego szeregu czasowego (tj. notowań indeksu SP500) przedstawia rysunek 5.2.

Rys. 5.2. ACF dla SP500

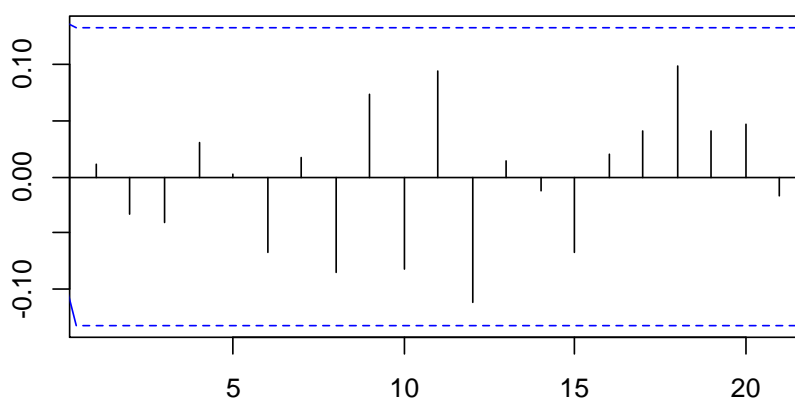


Z kolei rysunki 5.3 i 5.4 przedstawiają odpowiednio funkcję autokorelacji i autokorelacji cząstkowej dla szeregu czasowego pierwszych różnic notowań indeksu SP500.

Rys. 5.3. ACF dla pierwszych różnic



Rys. 5.4. PACF dla pierwszych różnic



Wiadomo także, że odchylenie standardowe wynosi:

- dla notowań: 306.8
- dla pierwszych różnic: 63.9.

Czy w oparciu o podane informacje można twierdzić, że przedstawiony na rysunku 5.1 szereg czasowy notowań indeksu SP500 jest realizacją procesu błędzenia losowego? Odpowiedź uzasadnij!

### Odpowiedzi

#### Odp. a)

W odpowiedzi należało wskazać, że szereg czasowy  $(y_t)_{t \in \mathbb{Z}}$  jest stacjonarny (słabo stacjonarny) jeżeli:

- wartość oczekiwana  $E(y_t)$  nie zależy od  $t$  ( $E(y_t) = \mu$ ,  $t \in \mathbb{Z}$ ),
- kowariancja między  $y_s$  a  $y_t$  zależy tylko od  $|s - t|$  ( $cov(y_s, y_t) = cov(y_{s+k}, y_{t+k})$ ,  $s, t, k \in \mathbb{Z}$ ).

Szczegóły np. w podrozdziale 7.3 w: "Regression Modeling with Actuarial and Financial Applications" - E.W. Frees, Cambridge, 2009.

**Odp. b)**

Proces błędzenia losowego definiuje się w następujący sposób:

$$y_t = y_{t-1} + c_t$$

gdzie  $c_t$  jest procesem białego szumu.

Szczegóły np. w podrozdziale 7.4 w: "Regression Modeling with Actuarial and Financial Applications" - E.W. Frees, Cambridge, 2009.

.....

**Odp. c)**

Na przykład sprawdzenie:

- czy w szeregu można wyodrębnić trend liniowy,
- czy wariancja  $y_t$  wzrasta wraz ze wzrostem  $t$ ,
- czy przyrosty  $y_t - y_{t-1}$  stanowią proces białego szumu,
- czy odchylenie standardowe szeregu przyrostów jest istotnie mniejsze w porównaniu z odchyleniem standardowym oryginalnego szeregu.

Należało wskazać i omówić co najmniej dwie metody.

Szczegóły np. w podrozdziale 7.4 w: "Regression Modeling with Actuarial and Financial Applications" - E.W. Frees, Cambridge, 2009.

.....

**Odp. d)**

Notowani indeksu SP500 są realizacją procesu błędzenia losowego.

Wskazuje na to:

- Funkcja autokorelacji szeregu czasowego notowań indeksu SP500. Dla procesu błędzenia losowego  $cov(y_t, y_s) = t\sigma_c^2, 1 \leq t \leq s$ .
- Funkcje autokorelacji i autokorelacji cząstkowej dla pierwszych różnic, które są charakterystyczne dla procesu białego szumu.
- Istotnie mniejsze odchylenie standardowe szeregu pierwszych różnic w porównaniu z odchyleniem standardowym oryginalnego szeregu.

---

**Rozwiązanie:**



**Zadanie 6.**

- a) (**2p**) Co rozumiemy przez pojęcia wariancja i obciążenie metody uczenia statystycznego?
- b) (**1p**) Rozważmy model  $Y = f(X) + \varepsilon$ ,  $X = (X_1, \dots, X_p)$ . Tutaj  $f$  jest pewną ustaloną, ale nieznaną funkcją  $X_1, \dots, X_p$ , a  $\varepsilon$  jest składnikiem losowym, który jest niezależny od  $X$  i ma średnią zero. Niech  $(y_0 - \hat{f}(x_0))^2$  oznacza kwadrat błędu predykcji dla obserwacji  $(x_0, y_0)$  ze zbioru testowego (czyli średni błąd kwadratowy (*mean squared error*, *MSE*) prognozy wyznaczonej dla obserwacji  $(x_0, y_0)$  ze zbioru testowego),  $\hat{f}$  jest oszacowaniem  $f$  na zbiorze uczącym. Wskaż (bez dowodu) na jakie trzy składowe można zdekomponować  $E[(y_0 - \hat{f}(x_0))^2]$ .
- c) (**2p**) Omów jak zmieniają się te składowe wraz ze zmianą elastyczności modelu, tj. jego zdolności do dopasowywania się do danych rzeczywistych (model charakteryzujący się większą elastycznością lepiej dopasowuje się do danych).

**Odpowiedzi:****Odp. a)**

W odpowiedzi należało wskazać, że:

- wariancja odnosi się do wielkości, o jaką zmieniają się  $\hat{f}$  otrzymane przy użyciu różnych zestawów danych uczących.
- obciążenie odnosi się do błędu, który jest wynikiem przybliżania rzeczywistego, bardzo skomplikowanego problemu znacznie prostszym modelem.

Szczegóły w podrozdziale 2.2 w “An Introduction to Statistical Learning with Applications in R” - G. James, D. Witten, T. Hastie, R. Tibshirani, Springer, 2021.

**Odp. b)**

W odpowiedzi należało wskazać, że:

$$E[(y_0 - \hat{f}(x_0))^2] = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\varepsilon)$$

Szczegóły w podrozdziale 2.2 w “An Introduction to Statistical Learning with Applications in R” - G. James, D. Witten, T. Hastie, R. Tibshirani, Springer, 2021.

**Odp. c)**

W odpowiedzi należało wskazać, że z reguły, wraz ze wzrostem elastyczności metod, wariancja wzrośnie, a obciążenie będzie się zmniejszało.

Szczegóły w podrozdziale 2.2 w “An Introduction to Statistical Learning with Applications in R” - G. James, D. Witten, T. Hastie, R. Tibshirani, Springer, 2021.

**Rozwiązanie:**

---

**Zadanie 7.**

- a) (1p.) Wymień co najmniej 4 czynniki, które należy wziąć pod uwagę podczas projektowania modelu.
- b) (2p.) Krótko opisz dwa czynniki spośród wymienionych w punkcie a).
- c) (2p.) Przedstaw wytyczne Krajowego Standardu Aktuarnego w zakresie jakości danych dotyczące:
  - walidacji danych,
  - braku danych.

**Odpowiedzi:**.....  
**Odp. a)**

Zobacz np. podrozdział 3.5.2.2 w: “Actuarial Aspects of ERM for Insurance Companies”, 2016.

.....  
**Odp. b)**

Zobacz np. podrozdział 3.5.2.2 w: “Actuarial Aspects of ERM for Insurance Companies”, 2016,

.....  
**Odp. c)**

Zobacz podrozdział 2.5 w: „Krajowy Standard Aktuarny Polskiego Stowarzyszenia Aktuariuszy – Praktyka Aktuarnia”, 2022,

**Zadanie 8.**

- a) (1p.) Krótko przedstaw ideę metod bootstrapowych.
- b) (2p.) Przedstaw algorytm postępowania w przypadku stosowania:
- nieparametrycznej metody bootstrapowej,
  - parametrycznej metody bootstrapowej.
- c) (2p.) Próba zawiera dwie obserwacje:  $x_1 = 2$  i  $x_2 = 4$ . Wykorzystując metodę bootstrapową oblicz błąd średniokwadratowy (*mean square error, MSE*) nieobciążonego estymatora średniej w populacji.

**Odpowiedzi:****Odp. a)**

Metody bootstrapowe należą do klasy metod symulacyjnych polegających na wnioskowaniu o interesującej nas wielkości na podstawie wielokrotnych replikacji oryginalnej próby. Przy czym replikacje uzyskuje się poprzez wielokrotne losowanie ze zwracaniem z próby (*bootstrap nieparametryczny*) lub założenie, że oryginalna próba pochodzi z ustalonej rodziny rozkładów, oszacowaniu jej parametrów (na podstawie oryginalnej próby), a następnie wylosowaniu z tego rozkładu replikacji (*bootstrap parametryczny*).

Szczegóły np. w podrozdziale 4.3 w: “Statistical Foundations of Actuarial Learning and its Applications” - M.V. Wuthrich, M. Merz, 2021.

**Odp. b)**

- Nieparametryczna metoda bootstrapowa – zobacz np. podrozdział 4.3.1 w: “Statistical Foundations of Actuarial Learning and its Applications” - M.V. Wuthrich, M. Merz, 2021.
- Parametryczna metoda bootstrapowa – zobacz np. podrozdział 4.3.2 w: “Statistical Foundations of Actuarial Learning and its Applications” - M.V. Wuthrich, M. Merz, 2021.

**Odp. c)**

Średnia z oryginalnej próby:  $\bar{x} = 3$

Próba	$\bar{x}_l$	$(\bar{x}_l - \bar{x})^2$
2, 2	2	1
2, 4	3	0
4, 2	3	0
4, 4	4	1
	$\Sigma$	2

$$\widehat{MSE} = \frac{2}{4} = 0.5$$

**Rozwiązanie:**

**Zadanie 9.**

- a) (**1p.**) Przedstaw ideę i konstrukcję wykresu prawdopodobieństwo-prawdopodobieństwo (*p-p plot, probability plot*). Wskaż zastosowanie tego wykresu.
- b) (**2p.**) Zanotowano następujące kwoty roszeń: 135, 29, 90, 64, 182. Dopasowano do nich rozkład wykładniczy, dla którego średnią oszacowano metodą największej wiarygodności. Skonstruuj i zinterpretuj wykres prawdopodobieństwo-prawdopodobieństwo.
- c) (**2p.**) Sprawdź dopasowanie tego rozkładu (tj. rozkładu z punktu b)) testem Kołmogorowa-Smirnowa. Przyjmij poziom istotności 0.05. Wartość krytyczna dla tego poziomu istotności wynosi:  $\frac{1.36}{\sqrt{n}}$ .

**Odpowiedzi:****Odp. a)**

Wykres prawdopodobieństwo-prawdopodobieństwo jest to jedna z graficznych metod oceny dopasowania modelu.

Konstrukcja:

- Wartości próbki porządkujemy w kolejności niemalejącej:  $x_1 \leq x_2 \leq \dots \leq x_n$
- W układzie współrzędnych zaznaczamy punkty o współrzędnych  $(\hat{F}_n(x_j), F(x_j)), j = 1, 2, \dots$ , gdzie  $\hat{F}_n(x_j)$  są wartościami dystrybuanty empirycznej ( $\hat{F}_n(x_j) = \frac{j}{n+1}$ ), a  $F(x_j)$  – wartościami dystrybuanty dopasowanego rozkładu.

Model jest dobrze dopasowany, jeśli punkty te leżą blisko linii  $y = x$ .

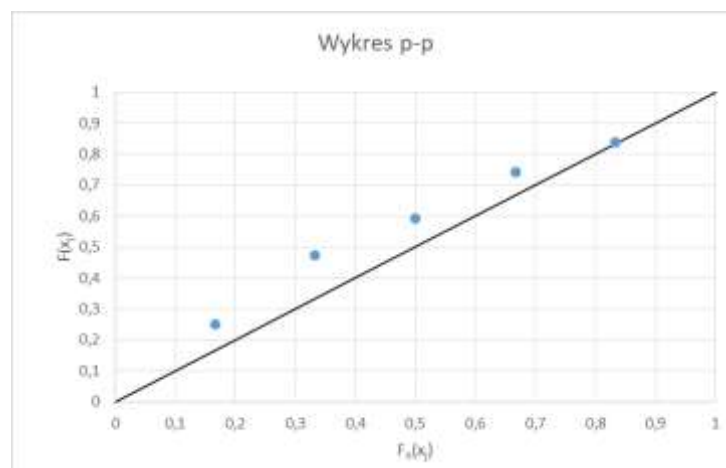
Szczegóły np. w podrozdziale 15.3 w: “Loss Models: From Data to Decisions”, 5th edition - S.A. Klugman, H.H Panjer, G.E. Willmot, Wiley, 2019.

**Odp. b)**

Dystrybuanta rozkładu wykładniczego:  $F(x) = 1 - e^{-\lambda x}, x \geq 0, E(X) = \frac{1}{\lambda}$

$$\hat{\lambda} = 0.01$$

$j$	$x_j$	$\hat{F}_n(x_j)$	$F(x_j)$
1	29	0.167	0.252
2	64	0.333	0.473
3	90	0.500	0.593
4	135	0.667	0.741
5	182	0.833	0.838



Na podstawie wykresu p-p możemy wstępnie przyjąć, że szkody podlegają rozkładowi wykładniczemu.

.....  
**Odp. c)**

Statystyka:  $D = \max_x |F_n(x) - F(x)|$ , gdzie  $F_n$  – dystrybuanta empiryczna,  $F$  – dystrybuanta dopasowanego rozkładu.

Obliczenia pomocnicze:

$j$	$x_j$	$F_n(x_j -)$	$F_n(x_j)$	$F(x_j)$	$ F_n(x_j -) - F(x_j) $	$ F_n(x_j) - F(x_j) $
1	29	0	0.2	0.252	0.252	0.052
2	64	0,2	0.4	0.473	0.273	0.073
3	90	0,4	0.6	0.593	0.193	0.007
4	135	0,6	0.8	0.741	0.141	0.059
5	182	0,8	1	0.838	0.038	0.162

Stąd otrzymujemy:  $D = 0.273$

Wartość krytyczna wynosi:  $\frac{1.36}{\sqrt{5}} = 0.608$ .

Wniosek: Nie ma podstaw do odrzucenia hipotezy, że szkody mają rozkład wykładniczy (na poziomie istotności 0.05).

Szczegóły np. w podrozdziale 15.4.1 w: “Loss Models: From Data to Decisions”, 5th edition - S.A. Klugman, H.H Panjer, G.E. Willmot, Wiley, 2019.

---

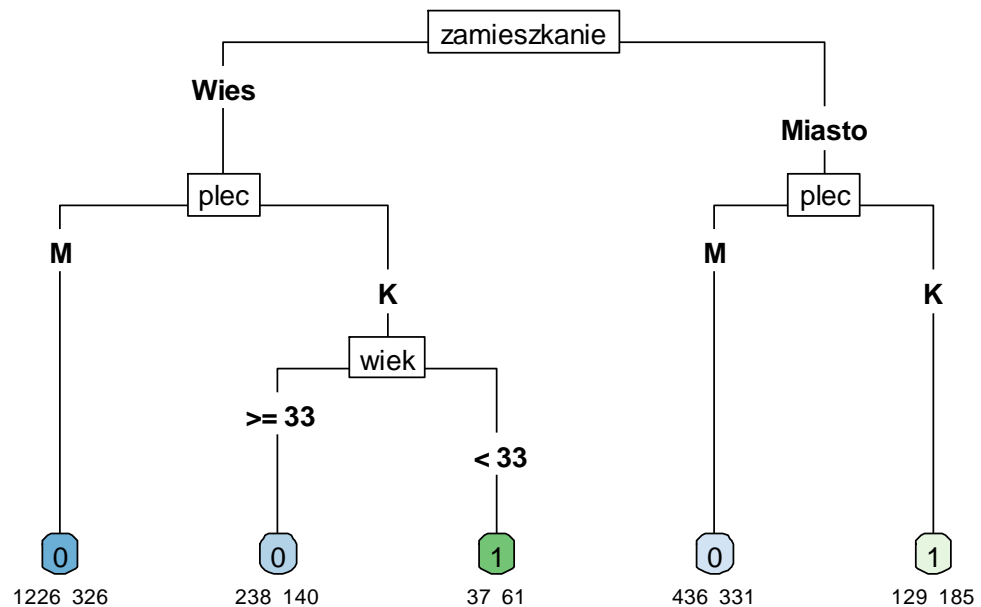
**Rozwiązanie:**

**Zadanie 10.**

- a) (2p.) Krótko przedstaw algorytm budowy binarnego drzewa klasyfikacyjnego.
- b) (2p.) Co to jest kryterium dobroci podziału (*goodness of split criterion*) wykorzystywane w konstrukcji takiego drzewa. Omów jedno wybrane kryterium dobroci podziału.
- c) (1p.) Przedłużenie umowy ubezpieczenia AC (zmienna  $Y_i$  przyjmująca dwie wartości:  $Y_i = 1$ , gdy kierowca przedłuży umowę na kolejny rok oraz  $Y_i = 0$ , gdy nie przedłuży) modelowano z wykorzystaniem następujących zmiennych objaśniających:
- *plec*: Płeć (K – kobieta, M – mężczyzna)
  - *zamieszkanie*: Miejsce zamieszkania (Miasto – miasto, Wies – wieś)
  - *wiek*: Wiek kierowcy w latach.

W tym celu skonstruowano następujące drzewo klasyfikacyjne (Rys. 10.1):

Rys. 10.1.



Na tym drzewie liczba z lewej strony pod liściem oznacza liczbę kierowców, którzy nie przedłużyli umowy, a liczba z prawej, którzy przedłużyli. Wykorzystując to drzewo, wyznacz prawdopodobieństwo przedłużenia umowy dla dwóch kobiet w wieku poniżej 33 lat, jednej mieszkającej na wsi a drugiej w mieście.

**Odpowiedzi****Odp. a)**

Zobacz np. w

- podrozdziale 8.1.2 w: “An Introduction to Statistical Learning with Applications in R” - G. James, D. Witten, T. Hastie, R. Tibshirani, Springer, 2021,

lub

- 
- podrozdziale 6.3 w: “Data Analytics for Non-Life Insurance Pricing” - M.V. Wüthrich, C. Buser, 2020.

---

**Odp. b)**

Zobacz np. w

- podrozdziale 8.1.2 w: “An Introduction to Statistical Learning with Applications in R” - G. James, D. Witten, T. Hastie, R. Tibshirani, Springer, 2021,

lub

- podrozdziale 6.3 w: “Data Analytics for Non-Life Insurance Pricing” - M.V. Wüthrich, C. Buser, 2020.

---

**Odp. c)**

Kobieta mieszkająca na wsi:  $\hat{p} = \frac{61}{37+61} = 0.622$

Kobieta mieszkająca w mieście:  $\hat{p} = \frac{185}{129+185} = 0.589$

---

**Rozwiązanie:**

---

**Sesja egzaminacyjna w dniu 24 stycznia 2023 r.****Modelowanie****Arkusz ocen**

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