

## 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 dewiancji

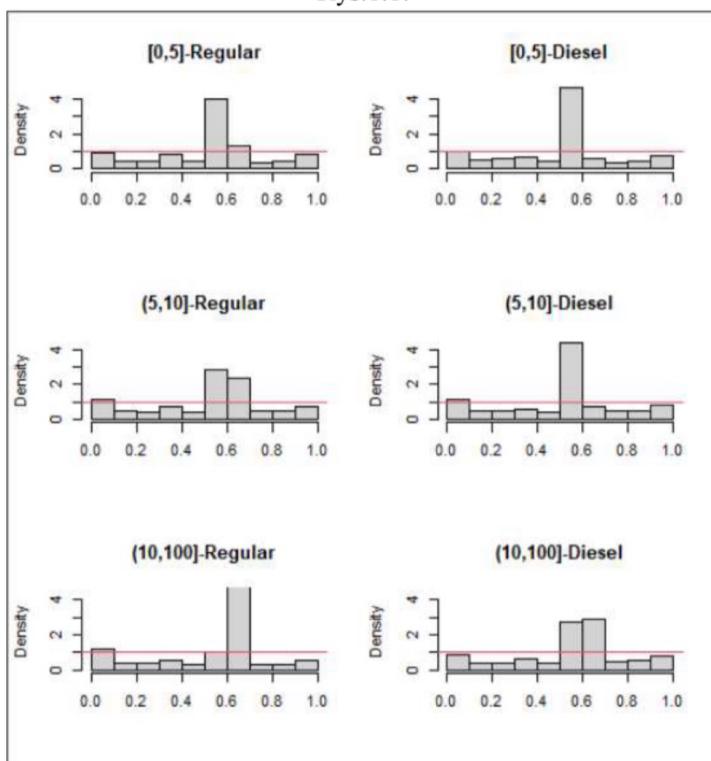
	Df	Deviance	Resid. Df	Resid. Dev	F	Pr(>F)
NULL		15866	11904			
CarAge.kat	2	25.886	15864	11878	16.564	6.511e-08 ***
Gas	1	11.024	15863	11867	14.108	0.0001732 ***
CarAge.kat:Gas	2	<b>14.606</b>	15861	11852	?	8.780e-05 ***

---

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

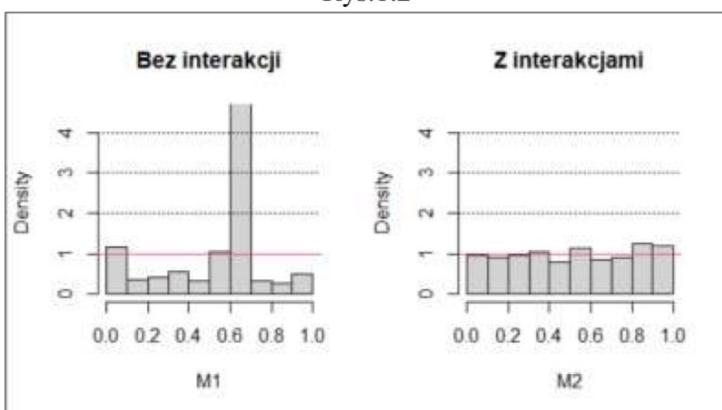
Wykorzystując zbiór testowy i model **M1**, dla każdej klasy ryzyka (tj. każdej kombinacji kategorii zmiennej *CarAge.kat* i *Gas*) skonstruowano histogram wartości  $F_{G_i}(y_{j,G_i})$ , gdzie:  $F_{G_i}$  – oszacowana dystrybuanta rozkładu gamma odpowiadająca klasie  $G_i$ ,  $y_{j,G_i}$  – wysokości 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? 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 roszczenia dla klasy: (10,100]-Regular.

a) Histogram dystrybuanty jest bardzo przydatne w ocenie jakości oszacowanych modeli. Histogram dystrybuanty to histogram wartości otrzymanych z tzw. transformaty całkowej prawdopodobieństwa (Probability Integral Transform - PIT). Jeżeli ciągła zmienna losowa  $X$  posiada dystrybuantę  $F$ , to  $F(X) \sim U(0, 1)$ , gdzie  $U(0, 1)$  oznacza rozkład jednostajny na przedziale  $(0, 1)$ . Oznacza to, że idealny histogram powinien być płaski – wszystkie słupki powinny mieć zbliżoną wysokość, oscylującą wokół czerwonej linii (gęstość równa 1).

b) Wybrano test F, ponieważ w testowanym modelu był szacowany parametr dyspersji.

Statystyka wynosi się wówczas:

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

gdzie:

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

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

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

Stąd:

$$F = \frac{14.606}{0.281385 \cdot 2} = 9.346225$$

c) Lepszy jest model M<sub>2</sub>, wskazuje na to wynik testu F: p-value ≤ 0.05.

Wynies diagnostyczny 1.2 wskazuje na model M<sub>2</sub>, stąd histogram mały nysokość ok. 1, wynies jest płaski.

d) Link kanoniczny dla rozkładu gamma to:

$$g(\mu) = \frac{1}{\mu}$$

$$\mu = \frac{1}{g(p)} \quad , \quad g(p) = \beta_0 + \beta_1 x_1 + \dots$$

$$y^p = \frac{1}{0.000759 + 0.00001922 + 0.000007362 + 0.0001143} = 110.44$$

## 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.

Q)

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.

b) Funkcja linowa:

$$g(\mu) = \ln \frac{\mu}{1-\mu}$$

$$e^g = \frac{\mu}{1-\mu}$$

$$e^g - \mu e^g - \mu = 0$$

$$\mu (1 + e^g) = e^g$$

$$\mu = \frac{\exp(g)}{1 + \exp(g)}$$

$$g = \beta_0 + \beta_1 X_1 + \dots, \quad \mu = p$$

Jedna z maksymalizować  $\exp(g)$  czyli mamy grupę klientów mieszkających  
w mieście z kategorii wiekowej G3.

Model A:

$$p = \frac{\exp[-4.272412 + 1.508494 + 0.043211 \cdot 100]}{1 + \exp[-4.272412 + 1.508494 + 0.043211 \cdot 100]} = 0.825952$$

Model B:

$$p = \frac{\exp[-3.627971 - 1.741291 + 0.034605 \cdot 100 + 0.0357 \cdot 100]}{1 + \exp[-3.627971 - 1.741291 + 0.034605 \cdot 100 + 0.0357 \cdot 100]} = 0.87666$$

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

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

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

Statystyka wyraża się wzorem:

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

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

$$\chi^2 = 23.2 > 5.991 = \chi^2_{0.05;2} \quad df = 2 \quad bo \quad 1859 - 1857 = 2$$

Odmawiamy  $H_0$  i przyjmujemy  $H_1$ .

### 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.) Aktuarusz 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 oczekiwana odpowiedniego rozkładu Poissona. Wyznacz logarytm funkcji wiarygodności dla tego modelu, a następnie znajdź oszacowania największej wiarygodności dla  $\alpha, \beta$  i  $\gamma$ .

$$a) f(y; \lambda) = \frac{\lambda^y e^{-\lambda}}{y!} = e^{\ln \lambda^y} \cdot e^{\ln e^{-\lambda}} e^{\ln(y!)^{-1}} = \exp\{y \ln \lambda - \lambda - \ln(y!)\}$$

stąd:

$$\Theta = \ln(\lambda)$$

$$\varphi = 1 \text{ czyli } a(\varphi) = 1$$

$$b(\Theta) = e^\Theta \quad \text{bo} \quad \Theta = \ln(\lambda) \Rightarrow \lambda = e^\Theta$$

$$c(y, \varphi) = -\ln(y!)$$

$$b) L(\lambda_A, \lambda_B, \lambda_C) = \prod_{i=1}^{35} \exp\{y_i \ln \lambda_i - \lambda_i - \ln(y_i!)\}$$

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

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

Obliczamy pochodne względowe:

$$\frac{\partial}{\partial \alpha} L = 11 - 10e^\alpha := 0$$

$$\frac{\partial}{\partial \beta} L = 3 - 5e^\beta := 0$$

$$\frac{\partial}{\partial x} L = 4 - 20e^x := 0$$

$$e^x = 1.1$$

$$e^y = 0.6$$

$$e^z = 0.2$$

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

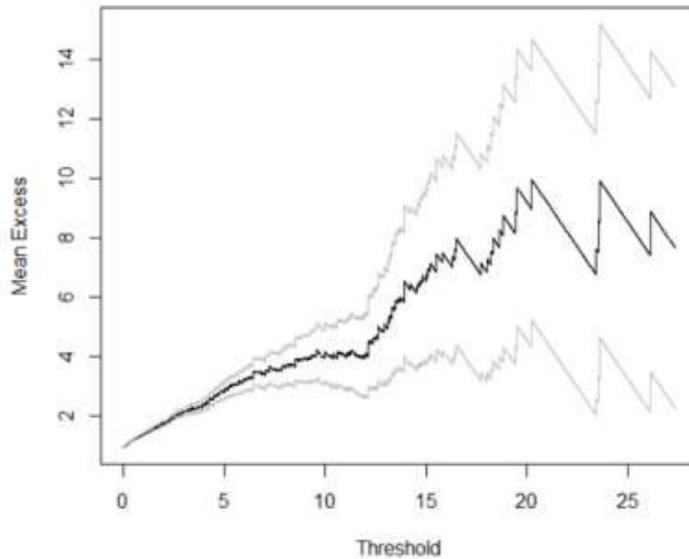
$$\hat{B} = \ln 0.6 = -0.51083$$

$$\hat{C} = \ln 0.2 = -1.38629$$

#### Zadanie 4.

- (1p.) Krótko przedstaw czym zajmuje się Teoria Wartości Ekstremalnych (*Extreme Value Theory, EVT*) i wskaż możliwości jej wykorzystania przez aktuariusza.
- (1p.) Jednym z głównych podejść wykorzystywanych w *EVT* jest analiza przekroczeń progu *POT* (*Peaks Over Threshold*). Krótko omów to podejście.
- (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*).
- (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.



a) Teoria Wartości Ekstremalnych to gałąź statystyki służąca do modelowania rzadkich, ekstremalnych zdarzeń, które znajdują się w "ogonach" rozkładów prawdopodobieństwa. *EVT* umożliwia prognozowanie prawdopodobieństwa wystąpienia rzadkich zdarzeń o wartościach wyższych niż kiedykolwiek zaobserwowane.

Wykorzystanie przez aktuariusza:

- Modelowanie śmiertelności w najstarszych latach w celu zamknięcia tablic trwania życia, co jest niezbędne przy kalkulacji rent dożywotnich.
- Obliczania wysokich kwantylów (*Value-at-Risk*) na potrzeby wymogów kapitałowych.

b) Podstawową ideą metody *POT* jest założenie, że jeśli interesuje nas ogon rozkładu (np. bardzo wysokie roszczenia ubezpieczeniowe), możemy wybrać wysoki próg  $u$  i analizować wszystkie wartości, które ten próg przekroczyły. Analizuje się nie same wartości, ale ich nadwyżki ponad ten próg, czyli wartości  $Y - u$ , pod warunkiem, że  $Y > u$ .

c)

Ustalenie progu:

- Zbyt niski próg: jeśli wybierzemy zbyt niski próg, uwzględnimy w analizie obserwacje, które w rzeczywistości nie należą do "ekstremalnego" ogona rozkładu. Włączenie danych spoza ogona prowadzi do tego, że model jest niedopasowany, a uzyskane estymaty parametrów są obciążone (błędne).
- Zbyt wysoki próg: spowoduje, że będziemy mieli bardzo mało danych (nadwyżek) do analizy. Mała próba danych prowadzi do estymacji parametrów o dużej wariancji.

Funkcja wartości oczekiwanej nadwyżki:

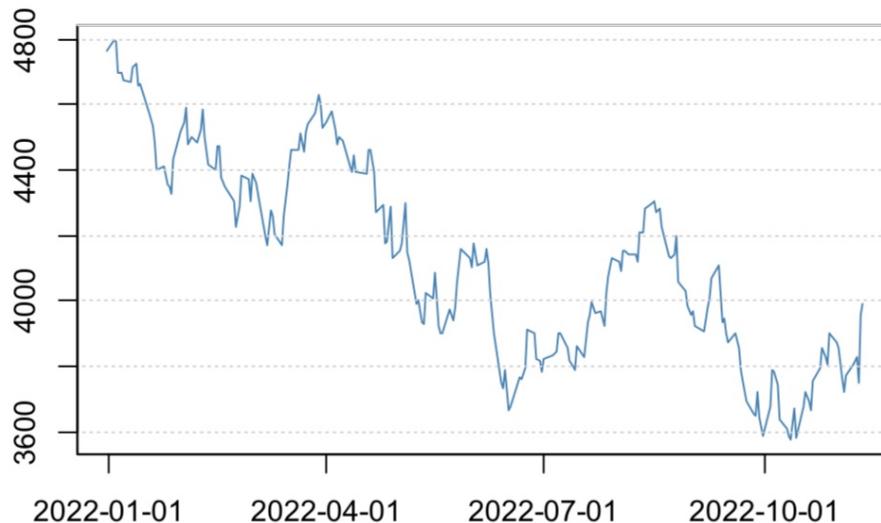
- ma liniową postać po przekroczeniu odpowiedniego progu  $u$ .

d) Po wyjęciu progu  $u = 12$  funkcja staje się liniowa.

## Zadanie 5.

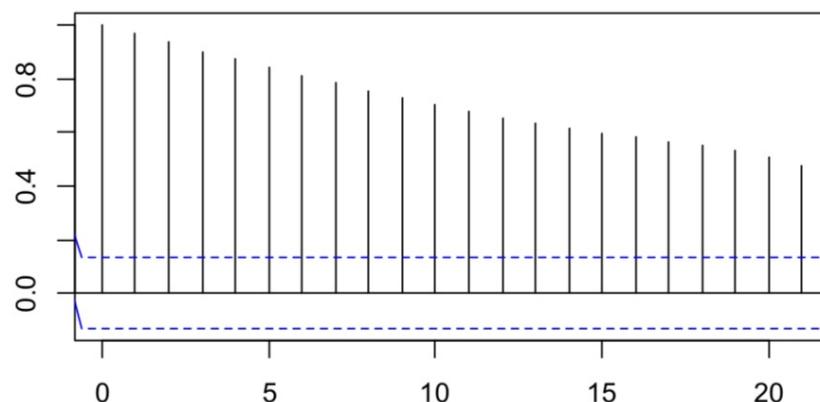
- a) (1p.) Wymień warunki stacjonarności (słabej) szeregu czasowego.
- b) (1p.) Podaj definicję procesu błądzenia losowego.
- c) (1p.) Wskaż i krótko przedstaw co najmniej dwie metody identyfikacji procesu błądzenia losowego.
- d) (2p.) Na poniższym wykresie (Rys. 5.1) podano notowania indeksu SP500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

Rys. 5.1. Notowania SP500



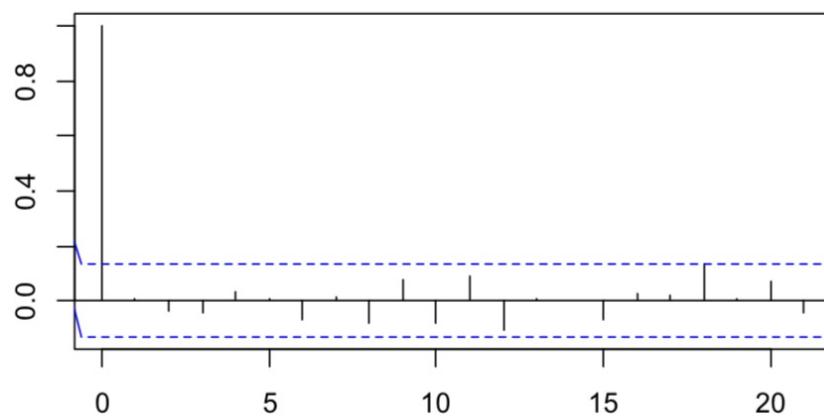
Funkcję autokorelacji dla tego szeregu czasowego (tj. notowań indeksu SP500) przedstawia rysunek 5.2.

Rys. 5.2. ACF dla SP500

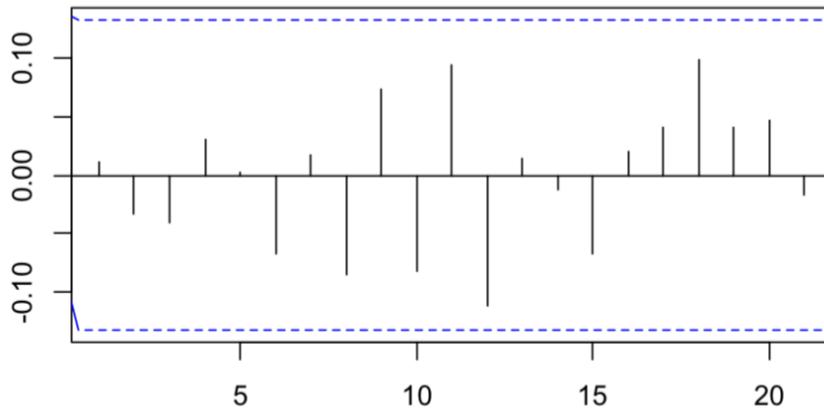


Z kolei rysunki 5.3 i 5.4 przedstawiają odpowiednio funkcję autokorelacji i autokorelacji częstotliwości dla szeregu czasowego pierwszych różnic notowań indeksu SP500.

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 szeregowy czasowy notowań indeksu SP500 jest realizacją procesu błędzenia losowego? Odpowiedź uzasadnij!

a) 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  $|s-t|$  ( $\text{cov}(y_s, y_t) = \text{cov}(y_{s+k}, y_{t+k}), s, t, k \in \mathbb{Z}$ ).

b) Proces błędzenia losowego definiuje się w następujący sposób:  $y_t = y_{t-1} + c_t$ , gdzie  $c_t$  jest procesem białego szumu.

c)

1. Czy przyrosty  $y_t - y_{t-1}$  stanowią proces białego szumu. Proces białego szumu jest stacjonarny i nie wykazuje żadnych widocznych wzorców w czasie. W praktyce polega to na stworzeniu nowego szeregu  $c_t = y_t - y_{t-1}$  i graficznej ocenie, czy jest on stacjonarny.
2. Czy odchylenie standardowe szeregu przyrostów jest istotnie mniejsze w porównaniu z odchyleniem standardowym oryginalnego szeregu.

d)

Tak, na podstawie dostarczonych informacji można z dużą pewnością twierdzić, że przedstawiony szeregowy czasowy notowań indeksu SP500 jest realizacją procesu błędzenia losowego.

Wnioski opierają się na kilku kluczowych obserwacjach, które są charakterystyczne dla procesu błędzenia losowego:

- Niestacjonarność oryginalnego szeregu: wykres notowań (Rys. 5.1) nie wykazuje stałej średniej, co jest typowe dla procesów niestacjonarnych. Potwierdza to funkcja autokorelacji (ACF) dla oryginalnego szeregu (Rys. 5.2), która pokazuje bardzo wysokie, dodatnie autokorelacje, powoli wygasające wraz ze wzrostem opóźnienia. Jest to klasyczny obraz ACF dla szeregu niestacjonarnego, w szczególności dla błędzenia losowego.

- Stacjonarność szeregu pierwszych różnic: kluczową cechą błądzenia losowego jest to, że jego pierwsze różnice tworzą stacjonarny proces białego szumu. Wykresy funkcji ACF (Rys. 5.3) i PACF (Rys. 5.4) dla pierwszych różnic potwierdzają tę właściwość. Niemal wszystkie słupki (poza lagiem 0 w ACF) mieszczą się wewnątrz niebieskich przedziałów ufności, co oznacza brak istotnej autokorelacji.
- Znacząca redukcja odchylenia standardowego: po zróżnicowaniu szeregu odchylenie standardowe spadło z 306,8 (dla notowań oryginalnych) do 63,9 (dla pierwszych różnic). Tak duża redukcja wariancji jest kolejnym silnym argumentem potwierdzającym, że mamy do czynienia z procesem błądzenia losowego.

## 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 zabiorze 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).

a)

1. Wariancja odnosi się do tego, jak bardzo oszacowanie funkcji  $\hat{f}$  zmieniłoby się, gdybyśmy je oszacowali na innym zbiorze danych treningowych. Mierzy ona wrażliwość modelu na niewielkie wahania w danych treningowych.
2. Obciążenie odnosi się do błędu, który jest wprowadzany przez przybliżenie bardzo skomplikowanego, rzeczywistego problemu za pomocą znacznie prostszego modelu. Innymi słowy, jest to błąd wynikający z założeń upraszczających, które przyjmujemy, aby ułatwić trenowanie funkcji docelowej.

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

c)

- Wariancja: rośnie wraz ze wzrostem elastyczności modelu. Bardziej elastyczne modele mocniej dopasowują się do danych treningowych, co oznacza, że niewielkie zmiany w tych danych mogą prowadzić do dużych zmian w oszacowanym modelu  $\hat{f}$ .
- Obciążenie: maleje wraz ze wzrostem elastyczności modelu. Mniej elastyczne modele (np. liniowe) opierają się na silnych, upraszczających założeniach co do prawdziwej postaci funkcji  $f$ , co wprowadza błąd systematyczny, czyli obciążenie. Bardziej elastyczne modele mają mniej założeń i mogą lepiej przybliżyć skomplikowane zależności, redukując obciążenie.
- Wariancja błędu losowego: ten składnik jest stały i nie zależy od elastyczności wybranego modelu. Wynika on z naturalnej losowości i szumu w danych, których nie da się wyeliminować, niezależnie od tego, jak dobry jest model. Stanowi on dolne ograniczenie dla błędu testowego.

## 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 Aktuarialnego w zakresie jakości danych dotyczące:
  - walidacji danych,
  - braku danych.

a) 1. Cel i proporcjonalność.

2. Czy model jest odpowiedni do mierzonych ryzyk.

3. Czas wykonania.

4. Ograniczenia.

- Czas wykonania: modele wewnętrzne mogą wiązać się z długim czasem wykonania, zwłaszcza jeśli stosowane są scenariusze stochastyczne. Dostępnych jest kilka technik pomagających skrócić czas wykonania, a aktuariusz powinien rozważyć i ocenić każdy potencjalny negatywny wpływ na dokładność. Przykłady technik skracających czas obejmują grupowanie danych w punkty modelowe, stosowanie scenariuszy deterministycznych zamiast stochastycznych dla portfeli bez opcji, równoważne rozwiązania analityczne (w postaci zamkniętej), zmniejszenie liczby scenariuszy stochastycznych oraz redukcję granularności czasowej. Ponadto, czasy wykonania można poprawić, stosując techniki redukcji wariancji podczas generowania scenariuszy stochastycznych, takie jak zmienne antytetyczne.
- Ograniczenia: modele zawsze będą miały ograniczenia statystyczne i teoretyczne. Nigdy nie można oczekiwać, że wyniki w pełni odwzorują świat rzeczywisty. Ważne jest, aby pamiętać o tych ograniczeniach podczas projektowania modelu i komunikowania jego wyników. Ważnym aspektem jest dokumentacja wszelkich istotnych ograniczeń w celu zapewnienia, że użytkownicy modelu są ich świadomi.

c) Walidacja danych: aktuariusz powinien podjąć uzasadnione kroki w celu sprawdzenia spójności, kompletności i dokładności wykorzystywanych danych. Możliwe kroki to między innymi:

- a. Uzgodnienie z audytowanymi sprawozdaniami finansowymi, zestawieniami obrotów i sald lub innymi stosownymi dokumentami, jeżeli są one dostępne;
- b. Przetestowanie danych pod kątem racjonalności w stosunku do zewnętrznych lub niezależnych danych;
- c. Przetestowanie danych pod kątem wewnętrznej spójności i spójności z innymi istotnymi informacjami; oraz
- d. Porównywanie danych z danymi za poprzedni okres lub okresy. Aktuariusz powinien opisać te kroki we wszystkich tworzonych raportach.

Brak danych: aktuariusz powinien wziąć pod uwagę możliwy wpływ wszelkich braków w danych (takich jak nieadekwatność, niespójność i niekompletność) na wyniki pracy. Braki, które prawdopodobnie nie będą miały istotnego wpływu na wyniki, nie muszą być dalej rozpatrywane. Jeżeli aktuariusz nie może znaleźć zadowalającego sposobu usunięcia braków, powinien rozważyć:

- a. Odmowę podjęcia lub kontynuowania świadczenia usług zawodowych;
- b. Współpracę ze zleceniodawcą w celu modyfikacji usług zawodowych lub uzyskania odpowiednich dodatkowych danych lub innych informacji; lub
- c. Z zastrzeżeniem zgodności z kodeksem etyki zawodowej, wykonanie usług zawodowych w najlepszy możliwy sposób i ujawnienie braków danych we wszystkich raportach (oraz wskazując potencjalny wpływ tych braków w danych).

## 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.

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).

b) Nieparametryczna metoda bootstrapowa:

1. Symulacja próby bootstrapowej: z oryginalnego zbioru danych  $(Y_1, \dots, Y_n)$  losujemy  $n$  obserwacji ze zwracaniem. Oznacza to, że każda obserwacja ma taką samą szansę na wylosowanie, a raz wylosowana obserwacja może zostać wylosowana ponownie. W ten sposób tworzymy nową, sztuczną próbę danych  $Y^*$ .
2. Obliczenie estymatora: na podstawie nowo utworzonej próbki bootstrapowej  $Y^*$  obliczamy interesujący nas estymator (np. średnią, wariancję, współczynnik regresji), stosując tę samą regułę decyzyjną  $A$ , co dla oryginalnej próbki. Otrzymujemy w ten sposób pojedynczą estymatę bootstrapową  $\hat{\theta}^* = A(Y^*)$ .
3. Powtórzenie: kroki 1 i 2 powtarzamy dużą liczbę razy, np.  $M = 1000$  lub więcej, uzyskując zbiór  $M$  estymat bootstrapowych  $(\hat{\theta}^{*(1)}, \dots, \hat{\theta}^{*(M)})$ .
4. Analiza wyników: otrzymany zbiór estymat tworzy empiryczny rozkład bootstrapowy. Na jego podstawie możemy oszacować właściwości pierwotnego estymatora  $\hat{\theta}$ , takie jak jego błąd standardowy, lub skonstruować dla niego przedziały ufności.

Parametryczna metoda bootstrapowa:

1. Estymacja parametrów: na podstawie oryginalnego zbioru danych  $(Y_1, \dots, Y_n)$  estymujemy nieznane parametry założonego rozkładu. Na przykład, jeśli zakładamy rozkład Poissona, estymujemy jego parametr  $\lambda$ . Otrzymujemy w ten sposób estymatę  $\hat{\theta}$ .
2. Symulacja próby bootstrapowej: generujemy nową, sztuczną próbę danych  $Y^*$  o wielkości  $n$ , losując obserwacje z dopasowanego rozkładu parametrycznego  $F(\cdot; \hat{\theta})$ . W przeciwieństwie do metody nieparametrycznej, nie losujemy tutaj z oryginalnych danych, lecz z rozkładu teoretycznego z wyestymowanymi parametrami.
3. Obliczenie estymatora: na podstawie nowo wygenerowanej próbki  $Y^*$  obliczamy interesujący nas estymator  $\hat{\theta}^* = A(Y^*)$ .
4. Powtórzenie: kroki 2 i 3 powtarzamy dużą liczbę razy ( $M$ ), uzyskując zbiór  $M$  estymat bootstrapowych  $(\hat{\theta}^{*(1)}, \dots, \hat{\theta}^{*(M)})$ .

5. Analiza wyników: podobnie jak w metodzie nieparametrycznej, otrzymany zbiór estymat tworzy empiryczny rozkład bootstrapowy, który służy do analizy właściwości pierwotnego estymatora  $\hat{\theta}$ .

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

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

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

## 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 roszczeń: 135, 29, 90, 64, 182. Dopasowano do nich rozkład wykładniczy, dla którego średnią oszacowano metodą największej wiarygodności. Skonstruj 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:  $\frac{1.36}{\sqrt{n}}$ .

### 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.

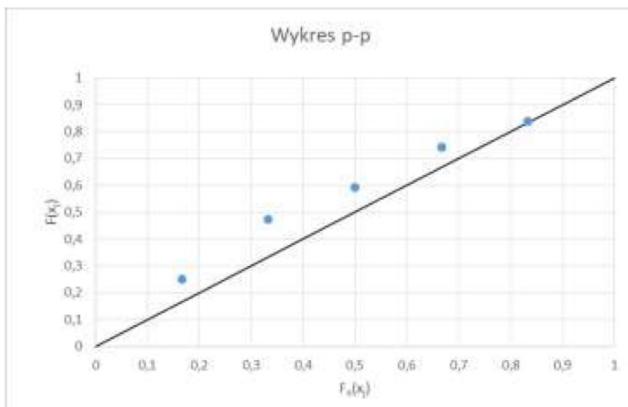
Model jest dobrze dopasowany, jeśli punkty te leżą blisko linii  $y = x$ .

### 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:  $\frac{1.36}{\sqrt{5}} = 0.608$ .

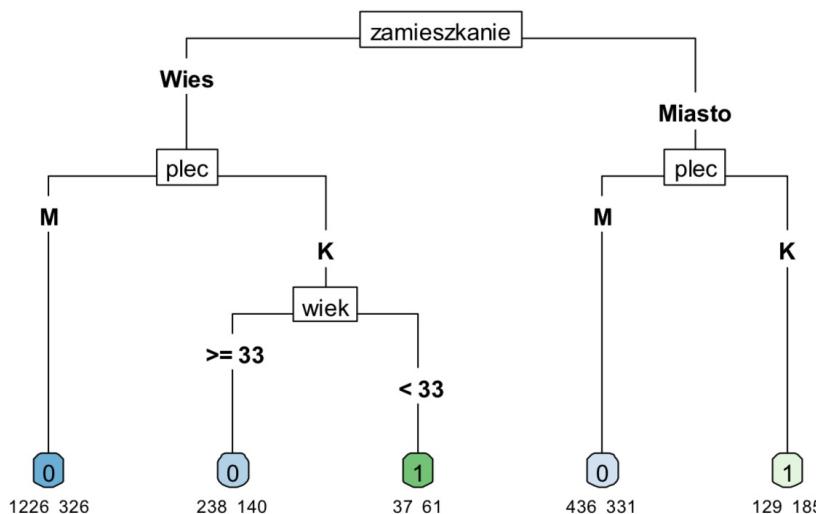
Wniosek: Nie ma podstaw do odrzucenia hipotezy, że szkody mają rozkład wykładniczy (na poziomie istotności 0.05).

## Zadanie 10.

- (2p.) Krótko przedstaw algorytm budowy binarnego drzewa klasyfikacyjnego.
- (2p.) Co to jest kryterium dobroci podziału (*goodness of split criterion*) wykorzystywane w konstrukcji takiego drzewa. Omów jedno wybrane kryterium dobroci podziału.
- (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.

Q)

1. **Start (korzeń drzewa):** Algorytm rozpoczyna się od jednego węzła (korzenia), który zawiera wszystkie obserwacje z zestawu treningowego.

2. **Znalezienie najlepszego podziału:**

- Algorytm iteracyjnie przeszukuje wszystkie predyktory (zmienne) i wszystkie możliwe punkty podziału dla każdego z nich.
- Dla każdego potencjalnego podziału obliczana jest miara "czystości" (lub "zanieczyszczenia") nowo powstałych węzłów potomnych. Najczęściej stosowane miary to:
  - **Indeks Giniego (Gini Index)**
  - **Entropia (Entropy)**
- Wybierany jest ten predyktor i ten punkt podziału, który prowadzi do największego zysku informacyjnego, czyli największej redukcji zanieczyszczenia w węzłach potomnych w porównaniu do węzła macierzystego.

3. **Podział (rekurencja):** Dane są dzielone na dwa nowe węzły (regiony) zgodnie z wybraną w kroku 2 regułą. Proces z kroku 2 jest następnie powtarzany rekurencyjnie dla każdego z nowo powstałych węzłów.

**4. Kryterium zatrzymania:** Proces dzielenia jest zatrzymywany, gdy spełniony zostanie jeden z warunków, np.:

- Węzeł jest idealnie "czysty" (zawiera obserwacje tylko jednej klasy).
- Osiągnięto maksymalną, zdefiniowaną wcześniej głębokość drzewa.
- Liczba obserwacji w węźle spadła poniżej określonego progu.

b)

To funkcja matematyczna używana w algorytmie budowy drzewa decyzyjnego do oceny, jak "dobry" jest każdy potencjalny podział węzła. Celem jest znalezienie takiego podziału (czyli takiej zmiennej i takiego jej progu), który w najlepszy sposób dzieli obserwacje na dwie grupy, które są jak najbardziej jednorodne pod względem klasyfikacji.

Indeks Giniego jest zdefiniowany wzorem:

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

jest to miara całkowitej wariancji we wszystkich  $K$  klasach. Nietrudno zauważyc, że indeks Giniego przyjmuje małą wartość, jeśli wszystkie  $\hat{p}_{mk}$  są bliskie zera lub jedynki. Z tego powodu indeks Giniego jest określany jako miara czystości węzła — mała wartość wskazuje, że węzeł zawiera w przeważającej mierze obserwacje z pojedynczej klasy.

c)

$$P_{m1} = \frac{61}{37+61} = 0,622$$

$$P_{m2} = \frac{185}{129+185} = 0,589$$