

## Zadanie 1.

- a) (2p.) Podaj definicje dewiancji i skalowanej dewiancji (*scaled deviance*). Wskaż, której wielkości, związanej z jakością modelu regresji liniowej, odpowiada dewiancja. Jaki rozkład ma skalowana dewiancja?
- b) (1p.) Twoim zadaniem jest wybór modelu o najlepszych zdolnościach predykcyjnych spośród zagnieźdzonych uogólnionych modeli liniowych. Wyjaśnij dlaczego podczas wyboru takiego modelu nie tylko dewiancja powinna zostać uwzględniona.
- c) (2p.) Oszacowano dwa zagnieźdzone uogólnione modele liniowe:  $M1$  i  $M2$ . W modelu  $M2$  w porównaniu z  $M1$  uwzględniono dodatkowo jakościową zmienną objaśniającą  $Z$ , przyjmującą wartości ze zbioru liczącego 6 kategorii. Zbiór uczący liczył 3000 obserwacji. Uzyskano między innymi następujące informacje:

	Model nasycony (Saturated Model)	Model M1	Model M2
Logarytm wiarygodności ( <i>Log-Likelihood</i> )	-700	-990	-985
Oszacowanie parametru dyspersji	2.0	2.0	2.0

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

4)

Dewiancja stanowi uogólnienie sumy kwadratów reszt i jest zdefiniowana jako:

$$D(\mathbf{y}, \hat{\mu}) = 2\phi(L_{full} - L(\hat{\beta}))$$

gdzie:

$\phi$  - parametr dyspersji. Dla niektórych rozkładów, takich jak rozkład Poissona czy dwumianowy, parametr ten jest stały i wynosi 1. Dla innych, jak rozkład Gamma, musi być estymowany.

$L_{full}$  - logarytm wiarygodności modelu pełnego (nasyconego). Jest to najwyższa możliwa wartość logarytmu wiarygodności, jaką można osiągnąć dla danych przy użyciu określonego rozkładu.

$L(\hat{\beta})$  - logarytm wiarygodności analizowanego modelu, obliczony dla estymowanych parametrów  $\hat{\beta}$ .

Dewiancja skalowana jest zdefiniowana jako:

$$\tilde{D}(\mathbf{y}, \hat{\mu}) = \frac{D(\mathbf{y}, \hat{\mu})}{\phi}$$

Dla dużych prób, rozkład dewiancji skalowanej można przybliżyć rozkładem chi-kwadrat ( $\chi^2$ ) z liczbą stopni swobody równą  $n - (p + 1)$ , gdzie  $n$  to liczba obserwacji, a  $p + 1$  to liczba szacowanych parametrów w modelu.

b)

Dzieje się tak, ponieważ dodanie kolejnych zmiennych do modelu prawie zawsze zmniejszy jego dewiancję (lub w najgorszym przypadku pozostawi ją bez zmian). Prowadzi to do ryzyka stworzenia modelu, który jest zbyt skomplikowany i świetnie dopasowuje się do danych uczących, ale traci zdolność do przewidywania nowych obserwacji.

c) Test ilorazu nieniygodności (test różnic dewiancji)

$H_0$ : dodatkowe parametry w modelu bardziej rozszerzonym są również zero

$H_1$ : dodatkowe parametry w modelu bardziej rozszerzonym nie są również zero

Statystyka testowa:

$$F = \frac{D(\hat{y}, \hat{\theta}^{M1}) - D(\hat{y}, \hat{\theta}^{M2})}{\hat{\sigma}^2_{q-p}} \sim F_{q-p, n-k-1}, \quad k+1 \text{ luba parametrow w bardziej rozszerzonego modelu}$$

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

$\hat{\sigma}^2$  - estymowana dyspersja modelu  $M2$

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

poniżej

$$D(\hat{y}, \hat{\theta}^{M1}) = 2(L_{\text{sat}} - L_{M1})$$

$$D(\hat{y}, \hat{\theta}^{M2}) = 2(L_{\text{sat}} - L_{M2}),$$

gdzie  $L_{\text{sat}}$  to nieniygodność modelu masywnego.

$q-p=5$  - zmienna  $Z$  ma 6 kategorii, więc dodanie jej do modelu wprowadza  $6-1=5$  dodatkowych parametrów (jedna kategoria staje się poziomem odmienienia)

$$\hat{\sigma}^2 = 2$$

Stosunek:

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

Wartość krytyczna:

$$F_{mn} = F_{q-p, m-n-1} = F_{5, \infty} = 2.216$$

$$F = 1 < 2.216 = F_{mn}$$

Pnajmniejszy  $H_0$  czyli model  $M_2$  nie jest lepszy.

Kryterium AIC:

$$AIC = -2L(\hat{\beta}) + 2p, \quad \text{gdzie } p \text{ to liczba parametrów.}$$

Model  $M_2$  ma  $p+5$  parametrów.

$$AIC_A = -2 \ln(L_{M_1}) + 2p = 2 \cdot 990 + 2p = 1980 + 2p$$

$$AIC_B = -2 \ln(L_{M_2}) + 2(p+5) = 2 \cdot 925 + 2p + 10 = 1950 + 2p$$

$$AIC_A = AIC_B$$

Na podstawie AIC modele są tak samo dobre.

## Zadanie 2.

- a) (2p.) Co to jest „*uporządkowana*” krzywa Lorenza (*ordered Lorenz curve*)?  
W jaki sposób może być wykorzystana w taryfikacji?
- b) (3p.) W procesie taryfikacji wykorzystywany jest model P1. Aktuariusz opracował inny konkurencyjny model P2. Modele te przetestował na zbiorze liczącym 5 obserwacji. Wyniki, które uzyskał przedstawia tabela 2.1. Skonstruuj „*uporządkowaną*” krzywą Lorenza. Krzywą przedstaw na rysunku 2.1 (w części Odp. b)). Opisz osie.

Tab. 2.1

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

a) Uporządkowana krzywa Lorenza to narzędzie graficzne służące do porównywania predykcyjnej mocy dwóch modeli taryfikacji składek ubezpieczeniowych, na przykład starego modelu z nowym. Pozwala ocenić, czy nowy model lepiej identyfikuje grupy ryzyka, które były niedokładnie wycenione w starym modelu, minimalizując w ten sposób ryzyko selekcji negatywnej. Podstawowym elementem jest pojęcie względności (relativity), czyli stosunku nowej składki do starej dla każdego klienta:

$$R = \frac{\hat{\mu}_2(X)}{\hat{\mu}_1(X)}$$

Główne zastosowanie polega na identyfikacji segmentów portfela, które są źle wycenione przez obecny model, co naraża ubezpieczyciela na selekcję negatywną.

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

$$R := \frac{1}{\hat{\mu}_1}$$

$$\frac{4,5}{5} = 0,9$$

$$\frac{7}{8} = 0,875$$

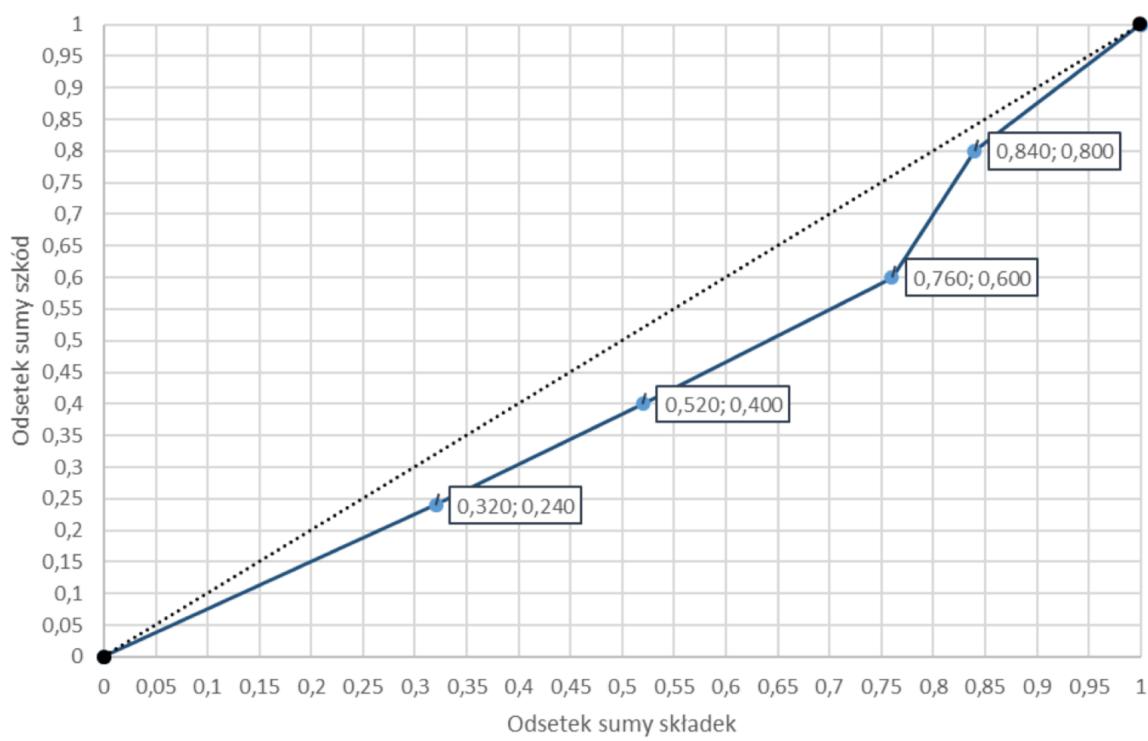
$$\frac{5,5}{6} = 0,916$$

$$\frac{2}{2} = 1$$

$$\frac{6}{4} = 1,5$$

Dane makiety uproszczonowej normowej według R:

Model $P_2$ ( $\hat{\mu}_2$ )	Model $P_1$ ( $\hat{\mu}_1$ )	Szkoły L	$A_i = \frac{\hat{\mu}_i}{\hat{\mu}_n}$	Skumulowane $\hat{\mu}_i$	Skumulowane L	Odcinek sumy składek $\hat{\mu}_i$	Odcinek sumy szkoł L
7	2	6	0,22	2	6	$\frac{2}{25} = 0,32$	$\frac{6}{25} = 0,24$
4,5	5	4	0,90	13	10	$\frac{13}{25} = 0,52$	$\frac{10}{25} = 0,40$
5,5	6	5	0,92	19	15	$\frac{19}{25} = 0,76$	$\frac{15}{25} = 0,60$
2	2	5	1,00	21	20	$\frac{21}{25} = 0,84$	$\frac{20}{25} = 0,80$
6	4	5	1,50	25	25	$\frac{25}{25} = 1$	$\frac{25}{25} = 1$



### Zadanie 3.

Liczب szkód (zmienna *L.szkod*) w pewnym portfelu ubezpieczeń AC modelowano z uwzględnieniem dwóch następujących zmiennych objaśniających:

*W.Kierowcy* – wiek kierowcy (zmienna jakościowa przyjmująca kategorie: W1, W2, W3),

*W.Samoch* – wiek samochodu (zmienna jakościowa przyjmująca kategorie: S1, S2).

Zebrano dane dotyczące liczby szkód zgłoszonych przez 67465 kierowców i przedstawiono je w tabeli 3.1 (w nawiasach podano ekspozycję na ryzyko).

Tab. 3.1

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

- a) (3p.) Twoim zadaniem jest oszacowanie, w oparciu o te dane, modelu regresji Poissona (z linkiem kanonicznym), z uwzględnieniem obydwu zmiennych objaśniających. Zakoduj zmienne objaśniające, przyjmując jako kategorie referencyjne: W1 i S1. Podaj:

- postać macierzy modelu (*model matrix*),
- wektor zmiennej zależnej (obserwacji)
- wektor przedstawiający ekspozycję na ryzyko.

- c) (2p.) Po oszacowaniu modelu dysponujesz następującymi wynikami:

- oszacowania parametrów

Coefficients:

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

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

- analiza dewiancji

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

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

## 2) Macierz modelu

(Intercept)	W.Kierowcy W2	W.Kierowcy W3	W.Samoch S2
1	0	0	0
1	1	0	0
1	0	1	0
1	0	0	1
1	1	0	1
1	0	1	1

Wektor zmiennej zakończenia (obserwacji)

L. skad

736
1315
179
782
1706
211

Wektor przedstawiający dysponicje na rynek

Dysponicja

3677.36
2312.54
1271.49
4225.76
1182.75
1828.04

b) Brakująca demografia:

$$26.446 - 3.236 = 23.21$$

Brakująca  $\Pr(>\text{Chi})$  - odczytane z tablicy:

$$\Pr(\chi^2 > 23.21) \leq 0.005 < 0.05 \quad - \text{lika stopni swobody} \rightarrow 3-2=1$$

Na tej podstawie na poziomie istotności 0.05 model z obejmującą zmiennymi objaśniającymi ma istotnie mniejszą demografię w porównaniu z modelem, w którym jedynie zmienne objaśniające jest niski hierarchyczny (W. Kierowcy).

Miary testu:

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

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

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 statystyki:

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

Stopnie swobody:  $3 - 2 = 1$

Wartość krytyczna na poziomie istotności 0.05 wynosi (z tablic): 3.841.

**Zadanie 4.**

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

Tab. 4.1

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

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

a) Aby wyznaczyć empiryczną funkcję przeżycia (ogivę), najpierw musimy obliczyć ile osób z połatkowej kohorty 100 000 osób dożywa do końca każdego z podanych przedziałów wiekowych.  $\hat{F}_{50}(t)$  - prawdopodobieństwo, że osoba w wieku 50 lat przeżyje co najmniej kolejne  $t$  lat.

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

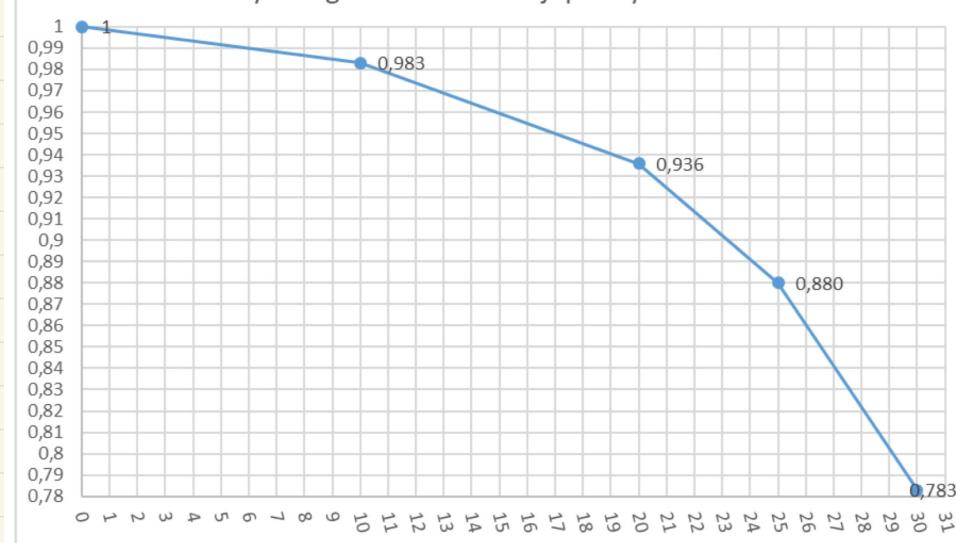
$$\hat{F}_{50}(10) = \frac{100\,000 - 1700}{100\,000} = 0,983$$

$$\hat{F}_{50}(20) = \frac{100\,000 - 1700 - 4700}{100\,000} = 0,936$$

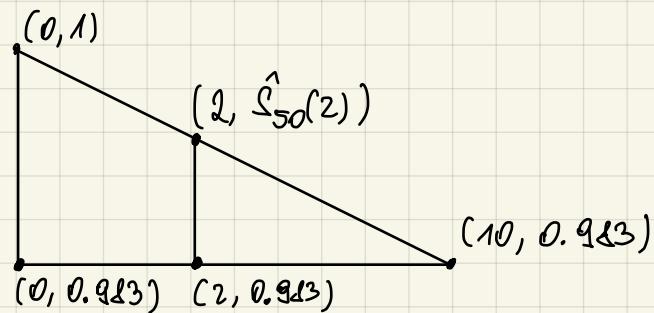
$$\hat{F}_{50}(25) = 0,880$$

$$\hat{F}_{50}(30) = 0,783$$

Krzywa ogiwowa dla funkcji przeżycia



b) Kompozycje z interpolacjami:



$$\frac{1 - 0.983}{10 - 0} = \frac{\hat{f}_{50}(2) - 0.983}{10 - 2}$$

$$\frac{0.983 - 0.936}{20 - 10} = \frac{\hat{f}_{50}(17) - 0.936}{20 - 17}$$

$$\frac{0.88 - 0.783}{30 - 25} = \frac{\hat{f}_{50}(28) - 0.783}{30 - 28}$$

$$\hat{f}_{50}(2) = 0.9966$$

$$\hat{f}_{50}(17) = 0.9501$$

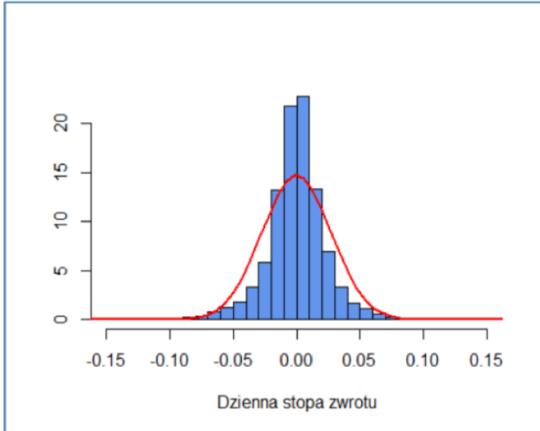
$$\hat{f}_{50}(28) = 0.8218$$

## Zadanie 5.

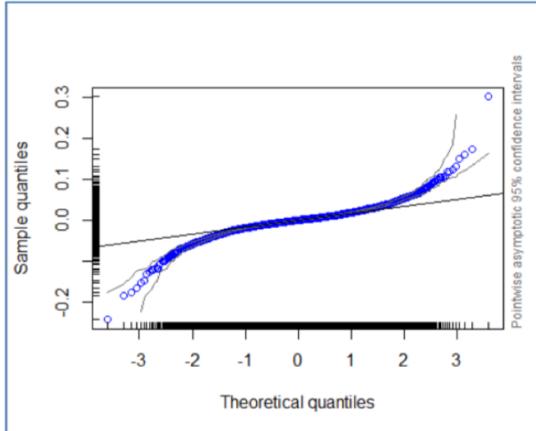
- a) (3p.) Wymień i krótko scharakteryzuj trzy wybrane właściwości finansowych szeregów czasowych (tzw. stylizowane fakty).
- b) (2p.) Na rysunku 5.1 przedstawiono 4 wykresy ilustrujące różne właściwości finansowego szeregu czasowego dziennych logarytmicznych stóp zwrotu spółki Aegon. Spośród wykresów A, B, C i D wybierz co najmniej trzy i przypisz im odpowiednią własność. (Uwaga, dwa punkty można uzyskać, gdy poprawnie zostaną zidentyfikowane co najmniej trzy stylizowane fakty, jeden - co najmniej dwa.)

Rys. 5.1

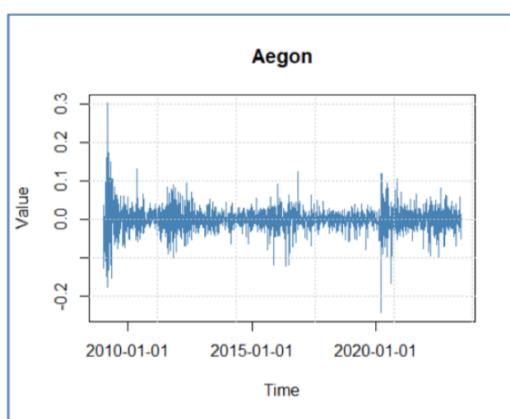
A. Histogram dla logarytmicznych stóp zwrotu i dopasowany rozkład normalny



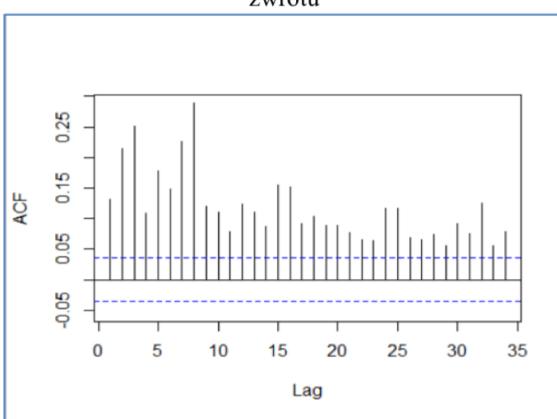
B. Wykres kwantyl-kwantyl logarytmicznych stóp zwrotu dla rozkładu normalnego.



C. Logarytmiczne stopy zwrotu



D. Autokorelacja modułów logarytmicznych stóp zwrotu



a)

1. **Grube ogony:** Rozkład stóp zwrotu nie jest rozkładem normalnym.

Charakteryzuje się tzw. "grubymi ogonami", co oznacza, że skrajne, bardzo wysokie lub bardzo niskie stopy zwrotu (gwałtowne wzrosty i spadki) występują znacznie częściej, niż przewidywałby to model oparty na rozkładzie normalnym. Jednocześnie obserwuje się wyższą koncentrację wyników wokół wartości średniej.

2. **Grupowanie się zmienności:** Zmienność stóp zwrotu nie jest stała w czasie. Obserwuje się okresy, w których wahania cen są niewielkie (niska zmienność), po których następują okresy o dużej amplitudzie wahań (wysoka zmienność). Innymi słowy, "dużym zmianom towarzyszą kolejne duże zmiany, a małym – małe".

3. **Brak autokorelacji stóp zwrotu, ale występowanie autokorelacji w wartościach bezwzględnych lub kwadratach stóp zwrotu:** Same stopy zwrotu są zazwyczaj nieskorelowane w czasie, co oznacza, że na podstawie przeszłych stóp zwrotu nie da się przewidzieć przyszłych. Jednak ich wartości bezwzględne (lub kwadraty), które są miarą zmienności, wykazują istotną, dodatnią i wolno wygasającą autokorelację. Jest to matematyczne potwierdzenie zjawiska grupowania się zmienności.

- b) A - histogram pokazuje, że dane nie pochodzą z rozkładu normalnego  
B - wykres pokazuje gubę ogony  
C - grupowanie zmienności  
D - dodatnia autokorrelacja modułów stop zwołu

**Zadanie 6.**

- a) (3p.) Na czym polega różnica między obserwacjami uciętymi (*truncated data*) a obserwacjami cenzurowanymi (*censored data*)? Wskaż i omów co najmniej jedną sytuację, w której aktuariusz może wykorzystywać:
- obserwacje ucięte,
  - obserwacje cenzurowane.
- b) (2p.) Wiadomo, że szkody w pewnym portfelu ubezpieczeń mają rozkład Weibulla z parametrem  $\tau = 2$ . W portfelu tym odnotowano 5 szkód. Wysokości trzech z nich były równe: 20, 30 i 45 tys. zł. O dwóch pozostałych wiadomo, że przekroczyły 50 tys. zł. Metodą największej wiarygodności oszacuj parametr  $\theta$  tego rozkładu.

Uwaga! Dla rozkładu Weibulla:

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

- a) Podstawowa różnica polega na ilości informacji dostępnej o obserwacji. W przypadku danych uciętych obserwacje spoza pewnego zakresu są całkowicie niewidoczne i nie są w ogóle rejestrowane. W przypadku danych cenzurowanych wiemy o istnieniu obserwacji, ale jej dokładna wartość jest nieznana.

#### Obserwacje ucięte (truncated data)

Obserwacja jest ucięta od dołu (left truncated) w punkcie  $d$ , jeśli jest rejestrowana tylko wtedy, gdy jej wartość jest większa od  $d$ . Jeżeli wartość jest równa lub mniejsza od  $d$ , obserwacja w ogóle nie jest zapisywana, a badacz nie wie o jej istnieniu.

##### i. Obserwacje ucięte

Aktuariusz analizuje dane dotyczące szkód komunikacyjnych. Polisy posiadają franszyzę integralną w wysokości  $d = 1000$  zł. Ubezpieczony, którego szkoda jest niższa niż 1000 zł, nie zgłasza jej, ponieważ nie otrzymałby odszkodowania.

##### ii. Obserwacje cenzurowane

Polisa ubezpieczeniowa pokrywa szkody do maksymalnej wysokości  $u = 100000$  zł. Jeśli rzeczywista szkoda jest wyższa (np. wynosi 150000 zł), ubezpieczyciel wypłaca 100000 zł i tylko ta informacja jest precyzyjnie rejestrowana.

- b) Funkcja prawdopodobieństwa:

$$f(x) = \frac{1}{x} \cdot \exp \left\{ -\left(\frac{x}{\theta}\right)^{\tau} \right\} = \exp \left\{ -\left(\frac{x}{\theta}\right)^{\tau} \right\}$$

#### Funkcja wiarygodności:

$$L(\theta) = f(20)f(30)f(45)[F(50)]^2$$

$$L(\theta) = \frac{2}{20} \left( \frac{20}{\theta} \right)^2 \exp \left\{ -\left( \frac{20}{\theta} \right)^2 \right\} \cdot \frac{2}{30} \left( \frac{30}{\theta} \right)^2 \exp \left\{ -\left( \frac{30}{\theta} \right)^2 \right\} \cdot$$

$$\cdot \frac{2}{45} \left( \frac{45}{\theta} \right)^2 \exp \left\{ -\left( \frac{45}{\theta} \right)^2 \right\} \cdot \exp \left\{ -2 \left( \frac{50}{\theta} \right)^2 \right\}$$

$$L(\theta) = \frac{2}{20 \cdot 30 \cdot 45} \cdot \frac{(20 \cdot 30 \cdot 45)^2}{\theta^6} \cdot \exp \left\{ -\frac{1325}{\theta^2} \right\}$$

$$L(\theta) := \theta^{-6} \cdot \exp \left\{ -\frac{1325}{\theta^2} \right\}$$

$$m(L(\theta)) = -6 \ln \theta - 1325 \theta^{-2}$$

$$m'(L(\theta)) = -6\theta^{-1} + 16650\theta^{-3}$$

$$16650\theta^{-3} = 6\theta^{-1} \mid \cdot \theta$$

$$16650\theta^{-2} = 6$$

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

### Zadanie 7.

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

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

a) *Boosting* to metoda uczenia zespołowego, która, podobnie jak *bagging*, łączy wiele prostych modeli, zwanych "słabyimi uczniami" (w tym kontekście drzewami decyzyjnymi), w jeden potężny model o wysokiej zdolności predykcyjnej.

Kluczową cechą tej metody jest to, że drzewa są budowane sekwencyjnie. Każde kolejne drzewo jest tworzone na podstawie informacji z drzew już istniejących w modelu. W przypadku problemu regresji, boosting działa w następujący sposób:

1. Dopasowuje się drzewo decyzyjne do danych.
2. Następnie kolejne drzewo jest dopasowywane nie do pierwotnej zmiennej odpowiedzi, ale do rezyduów ( błędów) z poprzedniego modelu.
3. Nowo utworzone drzewo jest dodawane do dopasowanej funkcji w celu aktualizacji rezyduów.

Dzięki temu podejściu model "uczy się powoli", stopniowo poprawiając swoje dopasowanie w obszarach, w których dotychczas działał słabo.

Metoda *boosting* jest podejściem ogólnym i może być stosowana zarówno w problemach regresji, jak i klasyfikacji.

b) *Bagging* tworzy wiele niezależnych drzew na losowych próbkach z oryginalnego zbioru danych, podczas gdy *boosting* buduje drzewa sekwencyjnie, gdzie każde kolejne drzewo uczy się na błędach poprzednich, wykorzystując zmodyfikowaną wersję oryginalnego zbioru danych.

- c)
1. Liczba drzew  $B$ : liczba ta mówi ile razy model jest ponownie trenowany, jeśli liczba drzew  $B$  jest zbyt duża model może zostać przetrenowany. Z tego powodu parametr ten dobiera się za pomocą cross-walidacji.
  2. Parametr kurczenia (shrinkage)  $\lambda$ : Jest to mała dodatnia liczba, która kontroluje tempo, w jakim *boosting* się uczy. Typowe wartości to 0.01 lub 0.001. Mniejsze wartości  $\lambda$  wymagają zazwyczaj większej liczby drzew  $B$ , aby osiągnąć dobrą wydajność.

## Zadanie 8.

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

b) (3p.) Liczbę szkód w pewnym portfelu ubezpieczeń AC modelowano z uwzględnieniem dwóch następujących zmiennych wyjaśniających:

*Plec* – płeć kierowcy (zmienna jakościowa: *K* (kobieta), *M* (mężczyzna)),

*Dystans* – przebyty dystans w ciągu roku (zmienna jakościowa, przyjmująca dwie kategorie: *poniżej 20 tys. km, powyżej. 20 tys. km*).

Zebrano dane dotyczące liczby szkód zgłoszonych przez 3000 kierowców i przedstawiono je w [tabeli 8.1](#) (w nawiasach podano ekspozycję na ryzyko):

Tab. 8.1

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

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

Uwaga! Wzór na współczynnik V Cramera:

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

Q) Współczynnik V Cramera jest miarą siły związku (asocjacji) między dwiema zmiennymi, często wykorzystywany w badaniach ubezpieczeniowych. Jego obliczenia opierają się na statystyce chi-kwadrat ( $\chi^2$ ) z tablicy kontyngencji, która zestawia ze sobą parę analizowanych cech. Jest on zdefiniowany wzorem:

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

gdzie:

- $\chi^2$  to wartość statystyki chi-kwadrat Pearsona dla testu niezależności.
- $n$  to liczba obserwacji.
- $k_1$  i  $k_2$  to liczba kategorii (poziomów) dla pierwszej i drugiej zmiennej.

Współczynnik V Craméra przyjmuje wartości z przedziału (0, 1).

Interpretacja tych wartości jest następująca:

- Wartość 0 - oznacza całkowitą niezależność między zmiennymi.
- Wartość 1 - oznacza idealną (pełną) zależność między zmiennymi.

Może być używany do oceny siły zależności dla różnych typów zmiennych, w tym:

- binarnych,
- kategorycznych (jakościowych),
- dyskretnych (skokowych).

Współczynnik V Craméra może być stosowany również dla zmiennych ciągłych (ilościowych), jednak wymaga to ich wcześniejszego przygotowania. Proces ten polega na dyskretyzacji (nazywanej również grupowaniem lub "bandingiem"), czyli podziale dziedziny zmiennej ciągłej na rozłączne przedziały. Po takim przekształceniu zmienna ilościowa jest traktowana jak zmienna dyskretna lub kategoryczna.

b)

$$\chi^2 = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

W tym przypadku :

$$E_{ij} = \frac{\text{Suma wiersza : Suma kolumny } j}{\text{Suma całkowita}}$$

	poniżej 20 tys. km	powyżej 20 tys. km	Suma
M	200	1800	2000
K	600	400	1000
Suma	800	2200	3000

Wartości teoretyczne :

	poniżej 20 tys. km	powyżej 20 tys. km
M	$\frac{400 \cdot 2000}{3000} = 533.333$	$\frac{2100 \cdot 2000}{3000} = 1466.667$
K	$\frac{400 \cdot 1000}{3000} = 266.667$	$\frac{2100 \cdot 1000}{3000} = 733.333$

$$\chi^2 = \frac{(200 - 533.333)^2}{533.333} + \frac{(1800 - 1466.667)^2}{1466.667} + \frac{(600 - 266.667)^2}{266.667} + \\ + \frac{(400 - 733.333)^2}{733.333} = 852.273$$

Wantoxic hipoteza to:

$$\chi^2_{(k_1-1)(k_2-1); 0.05} = \chi^2_{1; 0.05} = 3.841$$

$$\chi^2 = 852.273 > 3.841 = \chi^2_{1; 0.05}$$

Cyli zwiernej hipotezy: Dystans sa od siebie zalezne.

Wspolczynnik V Gramera:

$$V = \sqrt{\frac{152.273}{3000 \cdot \min\{2-1, 2-1\}}} = 0.533$$

Wspolczynnik V wskazuje na silne zaleznosc.

### Zadanie 9.

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

Tab. 9.1

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

Szkody są uporządkowane od największej do najmniejszej.

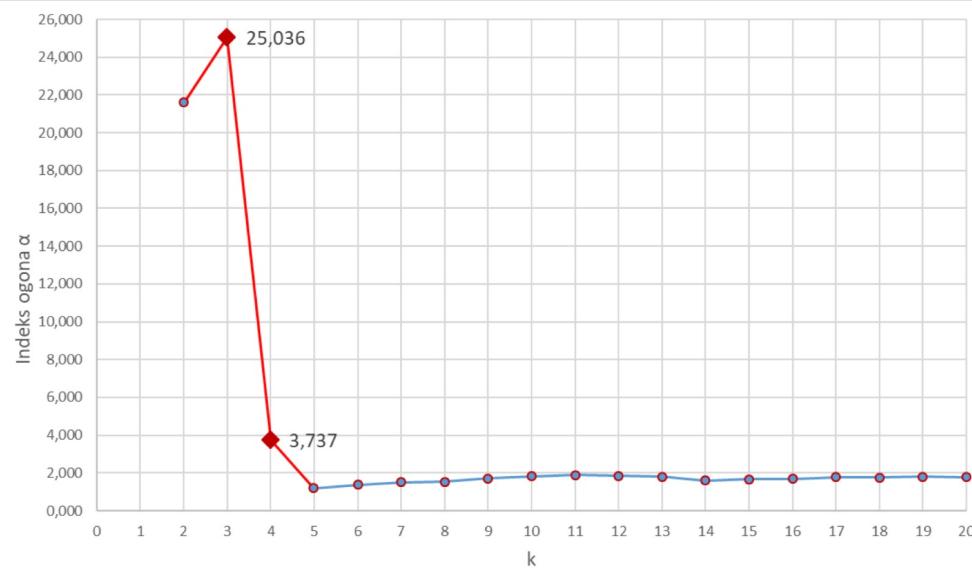
- (1p.) Co to jest estymator Hilla i w jaki sposób jest używany do analizy danych?
- (2p.) Na podstawie danych z tab. 9.1 skonstruowano wykres Hilla i przedstawiono go na rysunku 9.1 (część Odp. b)). Opisz osie wykresu i uzupełnij brakujący fragment.
- (2p.) Zaproponowano, aby wysokości szkód w tym portfelu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

Uwaga! Estymator Hilla ma postać:

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

a) Estymator Hilla to narzędzie statystyczne służące do estymacji (szacowania) indeksu ogona rozkładu prawdopodobieństwa. Indeks ogona, oznaczany jako  $\alpha$ , jest miarą "ciężkości" ogona rozkładu. Rozkłady o ciężkich ogonach charakteryzują się tym, że prawdopodobieństwo wystąpienia wartości ekstremalnych (bardzo dużych lub bardzo małych) jest znacznie wyższe niż w przypadku rozkładów o lekkich ogonach, takich jak rozkład normalny.

b) Rys. 9.1



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

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

$$= 25.036$$

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

c) Jaki jest to dystrybuanta rozkładu Pareto  $F(x) = 1 - \left(\frac{\theta}{x}\right)^\alpha$ ,  $x > \theta$ ,  
 w którym  $\theta = 1$ ,  $\alpha = 1.9$ . Parametr  $\alpha = 1.9$  mała warażnic  
 wykresem Hilla (stabilizuje się mniej więcej na poziomie 1.9).  
 Parametr  $\theta = 1$  to największa zaobserwowana wartość odniesiona  
 z tabeli.

### Zadanie 10.

- (2p.) Przedstaw indeks Giniego jako kryterium dobroci podziału (*goodness of split criterion*) w drzewach decyzyjnych.
- (3p.) Dysponujesz następującym (prostym) zbiorem danych dla zmiennych dychotomicznych (tab. 10.1):

Tab. 10.1

Zmienna zależna Y	Zmienne niezależne	
	X <sub>1</sub>	X <sub>2</sub>
1	Tak	B
1	Nie	A
0	Tak	B
0	Tak	A
1	Tak	A
0	Nie	B
0	Tak	B

Twoim zadaniem jest zbudowanie drzewa klasyfikacyjnego, w którym jako kryterium dobroci podziału wybrano indeks Giniego. Ustal, która zmienna powinna znajdować się w korzeniu drzewa (węźle głównym). Wiadomo, że w przypadku zmiennej X<sub>1</sub> średnia ważona odpowiednich indeksów Giniego wynosi 0.4857.

Uwaga! Indeks Giniego:

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

- a) 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.

- b) W zadaniu podany jest średni ważony indeks Giniego dla zm. X<sub>1</sub>, trzeba obliczyć to samo dla zm. X<sub>2</sub>, która dzieli zbior na dwie grupy :

$$X_1 = A, X_2 = B$$

Rozkład klasy Y dla podgrupy A :

$$Y = 0 : 1$$

$$Y = 1 : 2$$

$$p_0 = \frac{1}{3}$$

$$p_1 = \frac{2}{3}$$

Indeks Gimiego dla tej podgrupy:

$$\omega_A = p_0(1-p_0) + p_1(1-p_1) = \frac{1}{3} \cdot \frac{2}{3} + \frac{2}{3} \cdot \frac{1}{3} = \frac{4}{9}$$

Konkordancja Y dla podgrupy B:

$$Y=0 : 3$$

$$Y=1 : 1$$

$$p_0 = \frac{3}{4}$$

$$p_1 = \frac{1}{4}$$

$$\omega_B = \frac{3}{4} \cdot \frac{1}{4} + \frac{1}{4} \cdot \frac{3}{4} = \frac{3}{8}$$

Wagi:

$$\omega_A = \frac{3}{7} \quad \text{ponieważ w 3 tali obserwacje}$$

$$\omega_B = \frac{4}{7}$$

średnia ważona

$$\omega_{sw} = \frac{3}{7} \cdot \frac{4}{8} + \frac{4}{7} \cdot \frac{3}{8} = 0.405$$

Wybieramy mniejszą, która daje niższy średni ważony indeks Gimiego czyli m.  $X_2$ .