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 zagnieżdżonych uogólnionych modeli liniowych. Wyjaśnij dlaczego podczas wyboru takiego modelu nie tylko dewiancja powinna zostać uwzględniona.

c) ***(2p.)*** Oszacowano dwa zagnieżdżone 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 (*Log-Likelihood*)	-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*).

a) W kontekście uogólnionych modeli liniowych (GLM), dewiancja jest miarą jakości dopasowania modelu. Służy ona do oceny, jak dobrze dany model pasuje do obserwowanych danych.

Koncepcja Dewiancji

Dewiancja stanowi uogólnienie sumy kwadratów reszt, znanej z klasycznej regresji liniowej z rozkładem normalnym, na wszystkie modele z rodziny GLM. Kwantyfikuje ona zmienność w danych, która nie została wyjaśniona przez analizowany model.

Dewiancja:

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

gdzie: φ - parametr dyspersji (dla rozkładów Poissona lub dwumianowego wynosi 1)

L_{full} - logarytm wiarygodności modelu pełnego

$L(\hat{\beta})$ - logarytm wiarygodności modelu analizowanego

Dewiancja skalowana

$$\tilde{D}(y, \hat{\mu}) = \frac{1}{\varphi} D(y, \hat{\mu})$$

Dla dużych prób deviancja skalowana ma rozkład χ^2_{n-p-1}, gdzie n to liczba obserwacji, $p+1$ to liczba szacowanych parametrów w modelu.

b) Podczas wyboru najlepszego modelu predykcyjnego spośród zagnieżdżonych uogólnionych modeli liniowych, sama dewiancja nie jest wystarczającą miarą. 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 **przeuczenia** (ang. *overfitting*), czyli stworzenia modelu, który jest zbyt skomplikowany i świetnie dopasowuje się do danych uczących, ale traci zdolność do generalizacji i przewidywania nowych obserwacji.

Dlatego, aby dokonać właściwego wyboru, należy uwzględnić kompromis między **dobrocią dopasowania** (niską dewiancją) a **złożonością modelu** (liczbą parametrów). Służą do tego **kryteria informacyjne**, takie jak **kryterium informacyjne Akaikego (AIC)** lub **Bayesowskie kryterium informacyjne (BIC)**, które "karzą" model za posiadanie większej liczby parametrów. Model o niższej wartości kryterium informacyjnego jest uznawany za lepszy, ponieważ oferuje najlepszy kompromis między dopasowaniem a złożonością.

c) Test ilorazu wiarygodności (test różnicy deviancji)

H_0: dodatkowe parametry w modelu bardziej złożonym są równe zero

H_1: dodatkowe parametry w modelu bardziej złożonym nie są równe zero

Statystyka testowa:

$$F = \frac{D(y, \hat{\theta}^{M1}) - D(y, \hat{\theta}^{M1})}{\hat{\varphi}(q-p)} \sim F_{q-p,\, n-k-1}$$

, $k+1$ liczba parametrów bardziej złożonego modelu

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

$\hat{\varphi}$ – estymowana dyspersja modelu M2

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

ponieważ

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

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

gdzie L_{sat} to wiarygodność modelu nasyconego.

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

$\hat{\varphi}=2$

Stąd:

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

Wartość krytyczna:

$$F_{kr}=F_{q-p,\,n-k-1}=F_{5,\,3000-6}=F_{5,\infty}=2.216$$

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

Przyjmujemy H_0 czyli model M2 nie jest lepszy.

Kryterium AIC:

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

Model M2 ma $p+5$ parametrów.

$$AIC_A=-2\ln(L_{M1})+2p=2\cdot 990+2p=1980+2p$$

$$AIC_B=-2\ln(L_{M2})+2(p+5)=2\cdot 985+2p+10=1980+2p$$

$$AIC_A=AIC_B$$

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.

Uporządkowana krzywa Lorenza to wykres, który powstaje przez porównanie dwóch modeli predykcyjnych: starego ($\hat{\mu}_1$) i nowego ($\hat{\mu}_2$). Kluczowym elementem jest pojęcie **względności (relativity)**, czyli stosunku nowej składki do starej dla każdego klienta:

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

Niska wartość R oznacza, że nowy model proponuje znacznie niższą składkę niż stary, co sugeruje, że dany klient przepłacał w dotychczasowej taryfie.

Jak jest wykorzystywana w taryfikacji?

Uporządkowana krzywa Lorenza jest praktycznym narzędziem do oceny, czy nowy model taryfikacji ($\hat{\mu}_2$) przynosi poprawę w stosunku do starego modelu ($\hat{\mu}_1$).

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

b)

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

$R_i := \frac{\hat{\mu}_2}{\hat{\mu}_1}$

$\frac{4.5}{5} = 0.9$

$\frac{7}{8} = 0.875$

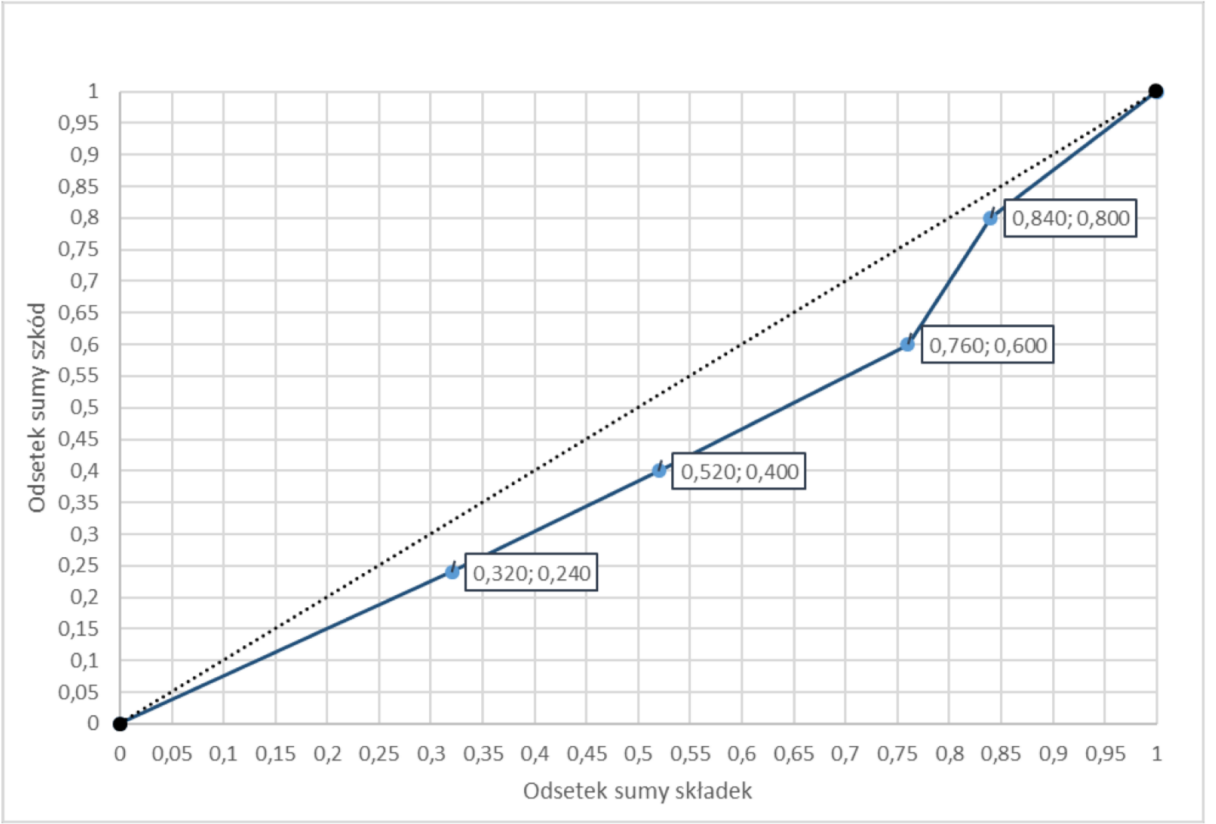
$\frac{5.5}{6} = 0.92$

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

$\frac{6}{4} = 1.5$

Dane należy uporządkować rosnąco według R_i:

Model P2 ($\hat{\mu}_2$)	Model P1 ($\hat{\mu}_1$)	Szkody L	$R_i = \frac{\hat{\mu}_2}{\hat{\mu}_1}$	Skumulowane $\hat{\mu}_1$	Skumulowane L	Odsetek sumy składek $\hat{\mu}_1$	Odsetek sumy szkód L
7	8	6	0.88	8	6	$\frac{8}{25} = 0.32$	$\frac{6}{25} = 0.24$
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.

Liczbę 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:

- <u>oszacowania parametrów</u>

```
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
```

- <u>analiza dewiancji</u>

```
           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 obydwoma zmiennymi objaśniającymi ma istotnie mniejszą dewiancję w porównaniu z modelem, w którym jedyną zmienną objaśniającą jest wiek kierowcy (*W.Kierowcy*)? Odpowiedź uzasadnij powołując się na wyniki odpowiedniego testu na poziomie istotności 0.05.

a) 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 zależnej (obserwacji)

L. szkod
736
1315
179
788
1706
211

Wektor przedstawiający ekspozycję na ryzyko

Ekspozycja
3677.36
8312.54
1271.49
4825.76
11882.75
1828.04

b) Bratująca deviancja:

$26.446 - 3.236 = 23.21$

Bratująca Pr(>Chi) - odczytane z tablicy:

$P(\chi_1^2 > 23.21) \leq 0.005 < 0.05$ - liczba stopni swobody to $3-2=1$

Na tej podstawie na poziomie istotności 0.05 model z obydwoma zmiennymi objaśniającymi ma istotnie mniejszą deviancję w porównaniu z modelem, w którym jedyną zmienną objaśniającą jest wiek kierowcy (W. Kierowcy).

Używając 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 statytstyki:

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

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

a) Aby wyznaczyć empiryczną funkcję przeżycia (ogiwę), najpierw musimy obliczyć ile osób z początkowej kohorty 100 000 osób dożyła do końca każdego z podanych przedziałów wiekowych. $\hat{S}_{50}(t)$ - p-stwo, że osoba w wieku 50 lat przeżyje co najmniej kolejne t lat.

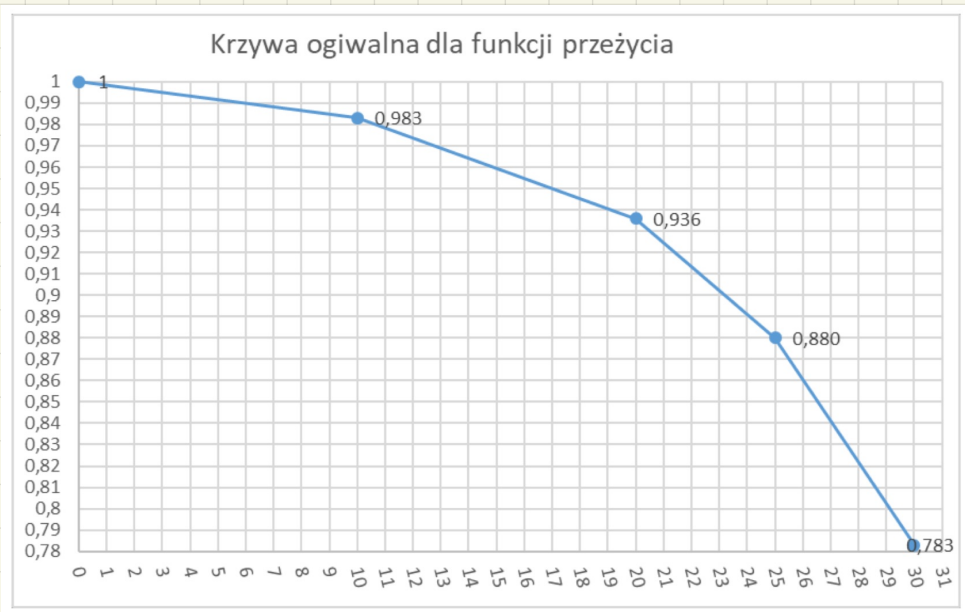
$\hat{S}_{50}(0) = 1$

$$\hat{S}_{50}(10) = \frac{100000 - 1700}{100000} = 0.983$$

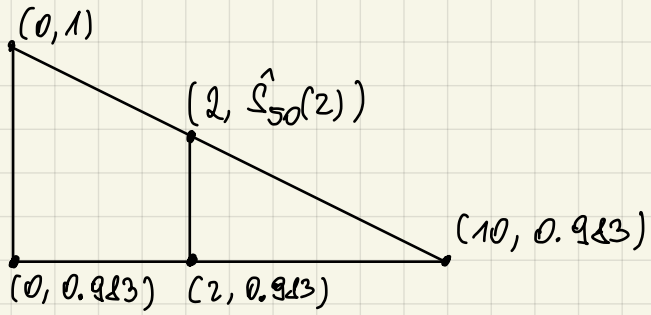
$$\hat{S}_{50}(20) = \frac{100\,000 - 1700 - 4700}{100\,000} = 0.936$$

$\hat{S}_{50}(25) = 0.880$

$\hat{S}_{50}(30) = 0.783$



b) Korzystając z interpolacji:



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

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

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

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

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

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

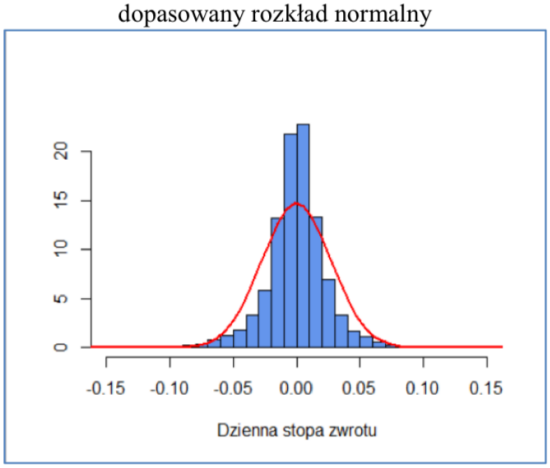
Zadanie 5.

a) (**3*p.***) Wymień i krótko scharakteryzuj trzy wybrane właściwości finansowych szeregów czasowych (tzw. stylizowane fakty).

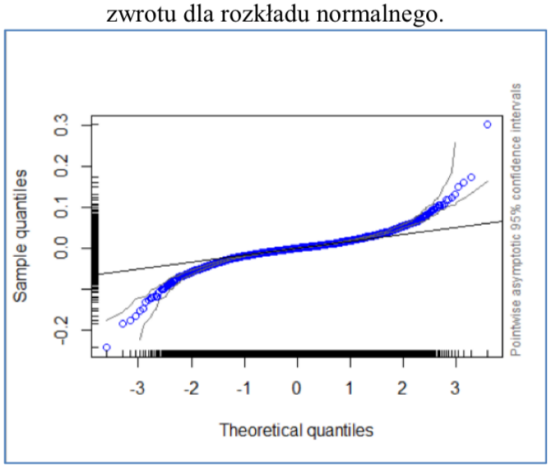
b) (**2*p.***) Na rysunku 5.1 przedstawiono 4 wykresy ilustrujące różne właściwości finansowego szeregu czasowego dziennych logarytmicznych stóp zwrotu spółki Aegon. Spośród wykresów A, B, C i D wybierz co najmniej trzy i przypisz im odpowiednią własność. (Uwaga, **dwa** punkty można uzyskać, gdy poprawnie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

Rys. 5.1

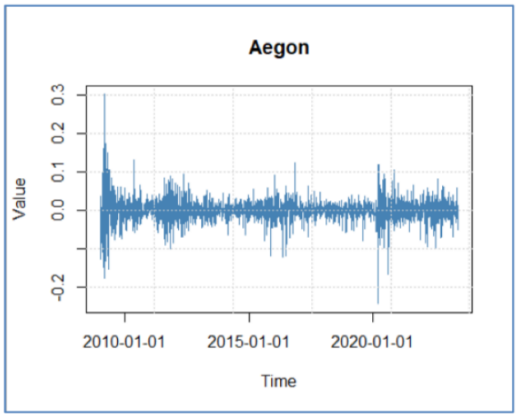
A. Histogram dla logarytmicznych stóp zwrotu i dopasowany rozkład normalny



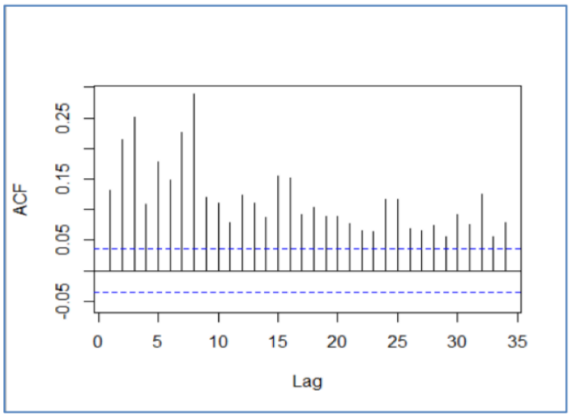
B. Wykres kwantyl-kwantyl logarytmicznych stóp zwrotu dla rozkładu normalnego.



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 grube ogony

C - grupowanie zmienności

D - dodatnia autokorelacja modułów stóp zwrotu

Zadanie 6.

a) (**3*p.***) 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:
   i. obserwacje ucięte,
   ii. obserwacje cenzurowane.

b) (**2*p.***) 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 θ tego rozkładu.

Uwaga! Dla rozkładu Weibulla:

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

a)

1. **Cenzurowanie z prawej strony** (ang. *right censored*) ma miejsce, gdy dla obserwacji osiągającej lub przekraczającej pewną wartość `u`, rejestruje się jedynie fakt, że wartość ta została osiągnięta, ale jej dokładna wielkość pozostaje nieznana. Obserwacje poniżej progu `u` są rejestrowane z ich dokładną wartością.

2. **Okrojenie z lewej strony** (ang. *left truncated*) występuje, gdy obserwacje poniżej pewnego progu `d` w ogóle nie są rejestrowane. W przypadku okrojenia nie wiedzielibyśmy nawet o istnieniu szkód, których wartość nie przekroczyła jakiegoś progu.

3. **Przykład dla danych cenzurowanych** Jeśli obserwacja ubezpieczonego kończy się przed jego śmiercią, jedyne, co wiemy, to to, że zgon następuje w pewnym momencie po czasie ostatniej obserwacji. Inną częstą sytuacją jest limit polisy, w przypadku którego, jeśli rzeczywista strata przekracza limit, wiadomo jedynie, że limit został przekroczony.

4. **Przykład dla danych uciętych** W przypadku polis sprzedanych przed rozpoczęciem okresu obserwacji, część ubezpieczonych umrze, podczas gdy inni dożyją, by rozpocząć obserwację. Nie tylko czasy ich zgonów nie zostaną zarejestrowane, ale nawet nie będziemy wiedzieć, ile ich było. Inną częstą sytuacją jest franszyza redukcyjna. Szkody poniżej franszyzy nie są rejestrowane i nie ma danych o tym, ile szkód było poniżej jej wartości.

b) Funkcja przeżycia:

$$S(x) = 1 - 1 + \exp\left\{-\left(\frac{x}{\theta}\right)^2\right\} = \exp\left\{-\left(\frac{x}{\theta}\right)^2\right\}$$

Funkcja wiarygodności:

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

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

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

$$L(\theta) = \frac{8}{20 \cdot 30 \cdot 45} \cdot \frac{(20 \cdot 30 \cdot 45)^2}{\theta^6} \cdot \exp\left\{-\frac{8325}{\theta^2}\right\}$$

$$L(\theta) := \theta^{-6} \cdot \exp\left\{-\frac{8325}{\theta^2}\right\}$$

$$\ln(L(\theta)) = -6\ln\theta - 8325\,\theta^{-2}$$

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

$$16650\,\theta^{-3} = 6\theta^{-1} \quad | \cdot \theta$$

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

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

Zadanie 7.
Jako aktuariusz wykorzystujesz modele bazujące na drzewach decyzyjnych. Chcesz skonstruować model o jak najlepszych zdolnościach predykcyjnych. Rozważasz możliwość wykorzystania metody uczenia zespołowego *boosting*.

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

Na czym polega metoda *boosting* i czy można ją stosować do problemów regresji i klasyfikacji?

Boosting (wzmacnianie) to metoda uczenia zespołowego (ang. *ensemble learning*), która polega na sekwencyjnym budowaniu modeli predykcyjnych, najczęściej drzew decyzyjnych. Każdy kolejny model jest trenowany w taki sposób, aby korygować błędy popełnione przez poprzednie modele. W przeciwieństwie do jednoczesnego budowania wielu niezależnych modeli (jak w metodzie *bagging*), *boosting* uczy się powoli, iteracyjnie poprawiając dopasowanie do danych.

Proces ten wygląda następująco:

1. Rozpoczyna się od dopasowania prostego modelu do danych.
2. Następnie analizuje się błędy (rezydua) tego modelu.
3. Kolejny model jest budowany tak, aby przewidywać te błędy (rezydua).
4. Nowy model jest dodawany do zespołu, a prognozy są aktualizowane. Rezydua są ponownie obliczane na podstawie prognoz całego, zaktualizowanego zespołu.
5. Kroki 3 i 4 są powtarzane wielokrotnie, co pozwala na stopniowe udoskonalanie modelu w obszarach, w których jego dotychczasowe działanie było najsłabsze.

Metoda **boosting może być stosowana zarówno do problemów regresji, jak i klasyfikacji**.

W jaki sposób metoda *boosting* różni się od metody *bagging* pod względem sposobu wykorzystania danych treningowych?

Główna różnica między metodami *boosting* i *bagging* polega na sposobie budowania modeli i wykorzystania danych treningowych:

- **Bagging (Bootstrap Aggregating):**
  - **Niezależne i równoległe budowanie modeli:** Każde drzewo decyzyjne jest budowane niezależnie od pozostałych.
  - **Próbkowanie bootstrapowe:** Każde drzewo jest trenowane na innej, losowej próbce danych treningowych, utworzonej przez losowanie ze zwracaniem (*bootstrap*). Oznacza to, że każda próbka treningowa jest nieco inna, ale pochodzi z tej samej oryginalnej puli danych.

- **Boosting:**
  - **Sekwencyjne i zależne budowanie modeli:** Drzewa są budowane jedno po drugim, a każde nowe drzewo jest tworzone z uwzględnieniem informacji o wynikach poprzednich drzew.
  - **Brak próbkowania bootstrapowego:** *Boosting* nie polega na losowaniu próbek ze zwracaniem. Zamiast tego, każde drzewo jest trenowane na zmodyfikowanej wersji oryginalnego zbioru danych.

Parametry dostrajania w metodzie *boosting*.

1. **Liczba drzew:** Jest to całkowita liczba drzew decyzyjnych, które zostaną sekwencyjnie zbudowane i włączone do modelu. W przeciwieństwie do *baggingu* i losowych lasów, zbyt duża liczba drzew w *boostingu* może prowadzić do przeuczenia (*overfittingu*), chociaż zjawisko to postępuje zazwyczaj powoli. Optymalną liczbę drzew często wybiera się przy użyciu walidacji krzyżowej.
2. **Współczynnik kurczenia (λ, ang. *shrinkage parameter* lub *learning rate*):** Jest to mała dodatnia liczba, która kontroluje tempo, w jakim *boosting* uczy się na błędach. Niższe wartości λ (np. 0.01 lub 0.001) spowalniają proces uczenia, co zazwyczaj wymaga większej liczby drzew do osiągnięcia dobrych wyników.
3. **Głębokość interakcji (d, ang. *interaction depth*):** Określa maksymalną liczbę podziałów w każdym drzewie, co kontroluje złożoność pojedynczego "słabego ucznia". Często dobrze sprawdza się wartość `d = 1`, co oznacza, że każde drzewo jest tzw. pniem decyzyjnym (ang. *stump*) z jednym podziałem.

Zadanie 8.

a) (**2*p*.**) 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) (**3*p*.**) Liczbę szkód w pewnym portfelu ubezpieczeń AC modelowano z uwzględnieniem dwóch następujących zmiennych objaśniających:
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

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

Wypowiedz się na temat zależności między zmiennymi objaśniającymi (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:

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

Współczynnik V Cramera jest miarą siły związku (asocjacji) między dwiema zmiennymi, opartą na statystyce chi-kwadrat Pearsona z tablicy kontyngencji.

Wartości i interpretacja

Współczynnik V Cramera przyjmuje wartości w przedziale od 0 do 1. Interpretacja jego wartości jest następująca:

- **V = 0** oznacza całkowity brak zależności statystycznej (niezależność) między analizowanymi zmiennymi.
- **V = 1** oznacza doskonałą zależność, gdzie znajomość wartości jednej zmiennej pozwala w pełni przewidzieć wartość drugiej.
- **Wartości pośrednie** wskazują na siłę związku – im wyższa wartość, tym silniejsza zależność.

Współczynnik ten jest tak znormalizowany, aby jego wartość była niezależna od liczby obserwacji oraz wymiarów tablicy kontyngencji. Oblicza się go za pomocą wzoru:

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

gdzie:

- χ^2 to statystyka chi-kwadrat Pearsona.
- n to całkowita liczba obserwacji.
- k_1 i k_2 to odpowiednio liczba kategorii (poziomów) dla pierwszej i drugiej zmiennej.

Zastosowanie i rodzaje zmiennych

Współczynnik V Cramera jest szeroko stosowany do oceny korelacji między cechami, zwłaszcza gdy klasyczne miary, takie jak współczynnik korelacji Pearsona, nie są odpowiednie dla danych nieciągłych.

Nie jest on ograniczony wyłącznie do zmiennych jakościowych (kategorycznych). Może być również stosowany do zmiennych ilościowych (ciągłych), jednak wymaga to ich uprzedniej **dyskretyzacji** (nazywanej również grupowaniem lub "bandingiem"). Proces ten polega na podzieleniu dziedziny zmiennej ciągłej na rozłączne przedziały, co w efekcie przekształca ją w zmienną kategoryczną. Po takiej transformacji współczynnik V Cramera oblicza się w standardowy sposób, jak dla zmiennych, które pierwotnie były jakościowe.

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

W tym przypadku:

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

	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	$\frac{800 \cdot 2000}{3000} = 533.333$	$\frac{2200 \cdot 2000}{3000} = 1466.667$
K	$\frac{800 \cdot 1000}{3000} = 266.667$	$\frac{2200 \cdot 1000}{3000} = 733.333$

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

$$+ \frac{(400 - 733.333)^2}{733.333} = 852.273$$

Wartość krytyczna to:

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

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

Czyli zmienne Płeć i Dystans są od siebie zależne.

Współczynnik V Cramera:

$$V = \sqrt{\frac{852.273}{3000 \cdot \min\{2-1, 2-1\}}} = 0.533$$

Współczynnik V wskazuje na silną zależność.

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.

a) ***(1p.)*** Co to jest estymator Hilla i w jaki sposób jest używany do analizy danych?
b) (***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.
c) (***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:

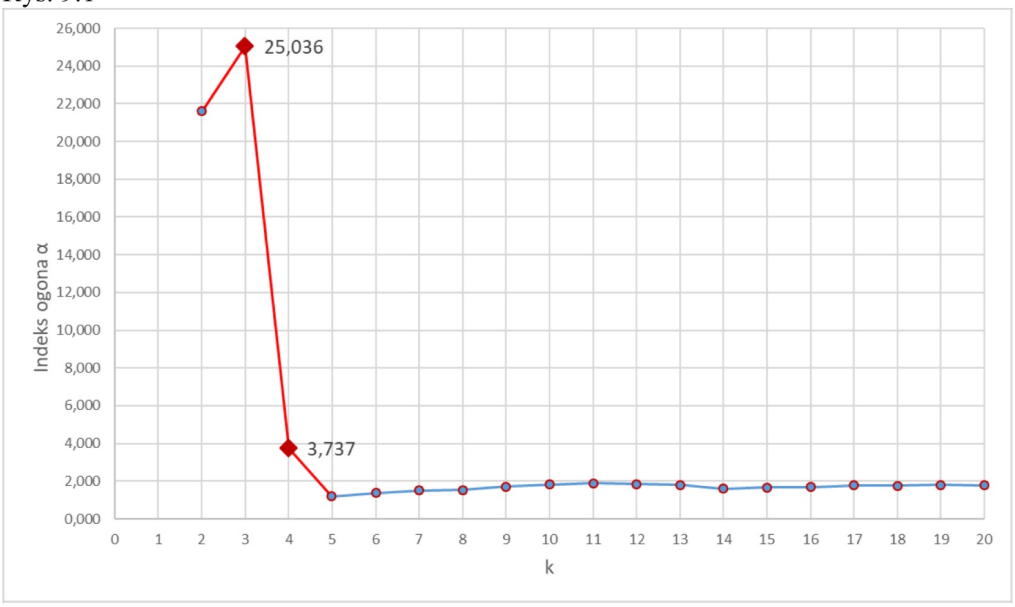
$$\hat{\alpha}_{k,n} = \left(\frac{1}{k}\sum_{j=1}^{k} \ln X_{j,n} - \ln X_{k,n}\right)^{-1}, 2 \le k \le n$$

a) Estymator Hilla to narzędzie statystyczne służące do estymacji (szacowania) indeksu ogona rozkładu prawdopodobieństwa. Indeks ogona, oznaczany jako α, 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.

Estymator Hilla znajduje zastosowanie w sytuacjach, gdzie kluczowe jest modelowanie szkód ekstremalnych (tzn. zdarzeń rzadkich, ale potencjalnie o dużej wartości). W ubezpieczeniach majątkowych jego użycie pozwala na lepsze zarządzanie ryzykiem i wycenę produktów ubezpieczeniowych. Jako przykład można wskazać:

- Wyznaczanie wysokości składki reasekuracyjnej. Firma ubezpieczeniowa zawiera umowę reasekuracyjną, aby zabezpieczyć się przed nadmiernymi stratami wynikającymi z ekstremalnych szkód (np. klęski żywiołowe). Estymator Hilla służy do oszacowania parametru grubości ogona rozkładu strat. Pozwala to na określenie prawdopodobieństwa wystąpienia szkody powyżej ustalonego progu.
- Modelowanie katastroficznych szkód majątkowych. W regionach zagrożonych katastrofami naturalnymi (huragany, powodzie) ubezpieczyciel musi oszacować wartość strat wynikających z rzadkich, ale ekstremalnych zdarzeń. Estymator Hilla pozwala na modelowanie ryzyka ekstremalnych szkód w oparciu o dane historyczne.

b) Rys. 9.1



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

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

$$= 25.036$$

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

c) Tak. Jest to dystrybuanta rozkładu Pareto $F(x) = 1 - \left(\frac{\theta}{x}\right)^{\alpha}$, $x > \theta$, w którym $\theta = 1$, $\alpha = 1.9$. Parametr $\alpha = 1.9$ można uzasadnić wykresem Hilla (stabilizuje się mniej więcej na poziomie 1.9). Parametr $\theta = 1$ to najmniejsza zaobserwowana szkoda odczytana z tabeli.

Zadanie 10.

a) (**2*p*.**) Przedstaw indeks Giniego jako kryterium dobroci podziału (*goodness of split criterion*) w drzewach decyzyjnych.

b) (**3*p*.**) Dysponujesz następującym (prostym) zbiorem danych dla zmiennych dychotomicznych (tab. 10.1):

Tab. 10.1

Zmienna zależna Y	Zmienne niezależne	
	X_1	X_2
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_1 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 miarą **zanieczyszczenia** (ang. *impurity*) lub niejednorodności w zbiorze danych. W kontekście drzew decyzyjnych służy jako kryterium do oceny, jak "dobry" jest potencjalny podział węzła na węzły potomne.

Główna zasada jest prosta: **dążymy do tworzenia podziałów, które skutkują jak najczystszymi węzłami potomnymi**. "Czysty" węzeł to taki, w którym większość (lub wszystkie) obserwacje należą do tej samej klasy.

1. **Wartość indeksu:** Indeks Giniego przyjmuje wartości od 0 do 0.5 (dla problemu klasyfikacji binarnej).
   - $G = 0$ oznacza **idealną czystość** – wszystkie obserwacje w węźle należą do jednej klasy.
   - $G = 0.5$ oznacza **maksymalne zanieczyszczenie** – obserwacje rozkładają się po równo między dwie klasy (np. 50% "Tak", 50% "Nie").
2. **Ocena podziału:** Dla każdej potencjalnej zmiennej, według której można dokonać podziału, algorytm oblicza średni ważony indeks Giniego dla węzłów potomnych, które by w wyniku tego podziału powstały.
3. **Wybór najlepszego podziału:** Wybierany jest taki podział (czyli taka zmienna), który prowadzi do **najniższej wartości średniego ważonego indeksu Giniego**. Oznacza to, że ten podział najlepiej separuje klasy i tworzy najczystsze możliwe podgrupy.

b) W zadaniu podany jest średni ważony indeks Giniego dla zm. X_1, trzeba obliczyć to samo dla zm. X_2, która dzieli zbiór na dwie grupy:

$X_2 = A$, $X_2 = B$

Rozkład klasy Y dla podgrupy A:

$Y = 0 : 1$

$Y = 1 : 2$

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

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

Indeks Giniego dla tej podgrupy:

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

Rozkład klasy Y dla podgrupy B:

$Y = 0 : 3$

$Y = 1 : 1$

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

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

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

Wagi:

$\omega_A = \frac{3}{7}$ ponieważ są 3 takie obserwacje

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

Średnia ważona

$G_{SW} = \frac{3}{7} \cdot \frac{4}{9} + \frac{4}{7} \cdot \frac{3}{8} = 0.405$

Wybieramy zmienną, która daje niższy średni ważony indeks Giniego czyli zm. X_2.