

Zadanie 1.

W tabeli 1.1 przedstawiono roczne dane dotyczące rezygnacji z odnowienia polisy w podziale na 3 segmenty taryfowe i 2 kanały sprzedaży. Pokazano stan na początku roku oraz liczbę rezygnacji w ciągu roku (ze stanu początkowego).

Tab. 1.1

Początek roku			Rezygnacje			
Segment	Kanał sprzedaży		Kanał sprzedaży			
	A	B	A	B		
	S1	60000	100000	3000	6000	
S2	100000	20000	2500	500		
S3	50000	400000	100	20000		

Wykorzystując przedstawione dane, modelowano prawdopodobieństwo rezygnacji dla indywidualnych umów w zależności od segmentu i kanału sprzedaży z wykorzystaniem uogólnionych modeli liniowych (GLM). Oszacowano dwa następujące modele:

Model M1:

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.95062	0.01710	-172.56	<2e-16 ***
SegmentS2	-0.74945	0.02339	-32.04	<2e-16 ***
SegmentS3	-0.20156	0.01454	-13.87	<2e-16 ***
KanalB	0.20220	0.01984	10.19	<2e-16 ***

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

Null deviance: 2094.257 on 5 degrees of freedom

Residual deviance: 86.814 on 2 degrees of freedom

AIC: 150.86

Model M2:

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.94444	0.01873	-157.190	< 2e-16 ***
SegmentS2	-0.71912	0.02759	-26.066	< 2e-16 ***
SegmentS3	-0.94738	0.10274	-9.221	< 2e-16 ***
KanalB	0.19290	0.02298	8.394	< 2e-16 ***
SegmentS2:KanalB	-0.19290	0.05468	-3.528	0.000419 ***
SegmentS3:KanalB	0.75448	0.10385	7.265	3.73e-13 ***

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

Null deviance: 2.0943e+03 on 5 degrees of freedom

Residual deviance: -6.3172e-11 on 0 degrees of freedom

AIC: 68.047

- (1p.) Wskaż jaki rozkład przyjęto dla zmiennej objaśnianej i podaj postać odpowiadającej jej kanonicznej funkcji łączącej (linku kanonicznego).
- (2p.) W oszacowanych modelach zastosowano kodowanie zero-jedynkowe. Wskaż kategorie referencyjne i podaj macierz modelu (*design matrix*) dla **M1** i dla **M2**.
- (2p.) Wykorzystując oba modele (**M1** i **M2**) oszacuj prawdopodobieństwo rezygnacji dla umowy z drugiego segmentu (S2) i kanału sprzedaży B. Porównaj wyniki z częstością rezygnacji dla tej grupy umów (tj. segment S2 i kanał B) wyznaczoną na podstawie danych (tab. 1.1). Skomentuj wynik porównania dla **M2** (tzn. wyjaśnij dlaczego uzyskano takie oszacowanie prawdopodobieństwa z wykorzystaniem tego modelu).

Q) Dla m. objaśnionej przyjęto wariancę zero-jedynkową.

Linię kanoniczną:

$$g(p_i) = \ln \frac{p_i}{1-p_i}, \text{ gdzie } p \text{ to prawdopodobieństwo rezygnacji}$$

b) Kategorie referencyjne:

- dla segmentu: SegmentS1

- dla kanału sprzedaży: KanalA

(braku ich na liście w odczycie dalego to kategorie referencyjne)

Macierz modelu dla M1:

(Intercept)	SegmentS2	SegmentS3	KanalB
1	0	0	0
1	0	0	1
1	1	0	0
1	1	0	1
1	0	1	0
1	0	1	1

Macierz modelu dla M2:

(Intercept)	SegmentS2	SegmentS3	KanalB	SegmentS2: KanalB	SegmentS3: KanalB
1	0	0	0	0	0
1	0	0	1	0	0
1	1	0	0	0	0
1	1	0	1	1	0
1	0	1	0	0	0
1	0	1	1	0	1

c) Model M1:

Wartość predykcyjna linijkowego:

$$\ln \frac{P_{S2,B}}{1 - P_{S2,B}} = -2.95062 - 0.74945 + 0.20220 = -3.497871$$

$$P_{S2,B} = 0.02937287$$

Model M2:

Wartość predykcyjna linijkowego:

$$\ln \frac{P_{S2,B}}{1 - P_{S2,B}} = -2.94444 - 0.71912 + 0.19290 - 0.19290 = -3.663562$$

$$\rho_{S2, B} = 0.025$$

Crestotęż regresji dla segmentu S2 i kanału B wynosi:

$$\frac{500}{20\,000} = 0.025$$

Prawdopodobieństwo szacowane za pomocą modelu M2 równa się crestotą, ponieważ jest to model masycony.

Liczba estymowanych parametrów w modelu (6) jest równa ilości unikalnych grup w danych ($3 \text{ segmenty} \cdot 2 \text{ kanały} = 6 \text{ grup}$). Model masycony ma wystarczającą elastyczność, aby idealnie dopasować się do danych, co skutkuje tym, że przedstawiane przez niego prawdopodobieństwa dla każdej grupy są dość dobrze równie obserwowanym w tych grupach crestotiom. Potwierdzeniem tego faktu w wynikach jest deviancja restowa bliska zero o 0 stopni swobody.

Zadanie 2.

- a) (2p.) Jakie są najważniejsze zalety i ograniczenia jądrowej estymacji funkcji gęstości w porównaniu z innymi technikami, takimi jak histogramy czy estymatory parametryczne?
- b) (3p.) Z pewnej polisy zbiorowej zanotowano następujące wypłaty (w tys. zł): 25, 30, 35, 35, 37, 39, 45, 47, 49, 55. Wykorzystując estymację jądrową z jądrem jednostajnym o stałej wygładzania 10, oszacuj prawdopodobieństwo wypłaty powyżej 40 tys. zł.

a) Zalety estymacji jądrowej funkcji gęstości:

- Metoda nieparametryczna, nie wymaga założenia konkretnego rozkładu, co pozwala na bardziej ogólną analizę danych.
- Za jej pomocą otrzymuje się gładką funkcję gęstości, która może lepiej odzwierciedlać rzeczywisty rozkład danych niż histogramy, szczególnie gdy dane są mało liczne lub mają skomplikowany rozkład.
- Dzięki stałej wygładzania (szerokości jądra) można kontrolować stopień wygładzenia estymowanej gęstości. Większa wartość stałej prowadzi do bardziej wygładzonej funkcji gęstości, podczas gdy mniejsza bardziej precyzyjnie odwzorowuje dane.
- Pozwala na porównywanie rozkładów różnych zestawów danych.

Ograniczenia jądrowej estymacji funkcji gęstości:

- Wymaga wyboru odpowiedniego jądra (np. gaussowskiego, Epanechikowa) oraz stałej wygładzania. Dobór tych parametrów może być subiektywny i wpływać na wyniki, a niewłaściwy ich wybór może prowadzić do błędnej estymacji rozkładu danych.
- Może być wymagająca obliczeniowo, szczególnie przy dużej ilości danych.
- Może niedokładnie odwzorować ogon rozkładu danych.

b) $\hat{F}(x) = \sum_{j=1}^k p(y_j) K_{y_j}(x)$, gdzie $p(y_j)$ to wiskotad empiryczny

$$K_y(x) = \begin{cases} 0 & x < y - b \\ \frac{x-y+b}{2b} & y - b \leq x \leq y + b \\ 1 & x > y + b \end{cases}$$

$$K_{25}(40) = 1 \quad \text{bo} \quad 40 > 25 + 10 = 35$$

$$K_{30}(40) = \frac{40-30+10}{20 \cdot 10} = 1$$

$$K_{35}(40) = 0.75$$

$$K_{37}(40) = 0.75$$

$$K_{39}(40) = 0.65$$

$$K_{41}(40) = 0.55$$

$$K_{45}(40) = 0.25$$

$$K_{47}(40) = 0.15$$

$$K_{49}(40) = 0.05$$

$$K_{55}(40) = 0 \quad \text{so} \quad 40 < 55 - 10 = 45$$

$$P(X_i) = \frac{1}{10}$$

$$\hat{F}(40) = \frac{1}{10}(1+1+0.75+0.75+0.65+0.55+0.25+0.15+0.05+0) = 0.515$$

$$P(X > 40) = 1 - P(X \leq 40) = 1 - \hat{F}(40) = 1 - 0.515 = 0.485$$

Zadanie 3.

- a) (3p.) Przedstaw koncepcję modelu DGLM (*Double Generalized Linear Model*) oraz krótko omów sposób estymacji jego parametrów.
b) (2p.) Wysokość pojedynczej szkody (zmienna *clm.incurred*) w pewnym portfelu ubezpieczeń AC modelowano z uwzględnieniem następujących zmiennych objaśniających:

ccm – pojemność silnika w cm³ (zmienna ilościowa),

nb.rb – zmienna jakościowa przyjmująca dwie kategorie:

NB – nowa polisa;

RB – wznowiona polisa,

driver.gender – płeć kierowcy (zmienna jakościowa: *Female*, *Male*).

Wstępna analiza danych wykazała, że wariancja wysokości pojedynczej szkody dla nowych polis była ponad półtora raza większa od wariancji dla polis wznowionych. W związku z tym, na podstawie zbioru uczącego oszacowano *Double Generalized Linear Model*, w którym przyjęto rozkład gamma dla zmiennej objaśnianej oraz link logarytmiczny dla obydwu modeli. Uzyskano następujące wyniki:

Call: dglm(formula = clm.incurred ~ ccm + nb.rb + driver.gender, dformula = ~nb.rb + ccm, family = Gamma(link = "log"), data = zbior.uczacy)

Mean Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.906079e+00	1.201117e-01	57.4971318	0.0000000000
ccm	4.048551e-05	6.569563e-05	0.6162588	0.537768064
nb.rbRB	-1.907909e-01	6.412235e-02	-2.9754202	0.002948009
driver.genderMale	-2.085674e-01	6.370910e-02	-3.2737456	0.001072737

(Dispersion Parameters for Gamma family estimated as below)

Scaled Null Deviance: 3942.734 on 3168 degrees of freedom

Scaled Residual Deviance: 3918.855 on 3165 degrees of freedom

Dispersion Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	3.768577e-01	9.429130e-02	3.996739	6.422112e-05
nb.rbRB	-6.159272e-02	4.975300e-02	-1.237970	2.157273e-01
ccm	9.304519e-05	5.556936e-05	1.674397	9.405254e-02
(Dispersion parameter for Digamma family taken to be 2)				

Scaled Null Deviance: 3632.966 on 3168 degrees of freedom

Scaled Residual Deviance: 3710.548 on 3166 degrees of freedom

Minus Twice the Log-Likelihood: 48391.39

Number of Alternating Iterations: 5

Wiadomo, że podstawowe statystki opisowe dla zmiennej *ccm* w zbiorze uczącym są następujące:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
970	1398	1560	1643	1868	2700

Na podstawie podanych wyników wyznacz:

- możliwe minimalne i możliwe maksymalne oszacowanie parametru dyspersji dla polis w zbiorze uczącym,
- oszacowania wartości oczekiwanej i wariancji zmiennej losowej modelującej wysokość pojedynczej szkody dla nowej polisy wystawionej na samochód o pojemności 2700 cm³, którego właścicielem jest mężczyzna.

q)

Proces estymacji podwójnego uogólnionego modelu liniowego (DGLM) jest **procesem iteracyjnym**, który naprzemiennie dopasowuje model dla wartości średniej i model dla dyspersji, aż do osiągnięcia zbieżności. Modele te "współdziałają" ze sobą, wymieniając się informacjami na każdym etapie.

Kluczowe etapy procesu

1. Dopasuj GLM dla wartości średniej, ze stałym ϕ dla wszystkich obserwacji.
2. Oblicz wkład każdej obserwacji do dewiacji i oblicz kwadrat Pearsona lub dewiancję residiów R_i^2 .
3. Dopasuj GLM dla dyspersji, przyjmując jako zmienną objaśnianą kwadraty residiów R_i^2 . Przyjmuje się rozkład Gamma i na tym etapie nie uwzględnia się wag. Dopasowane wartości stają się nowym parametrem dyspersji dla każdej obserwacji.
4. Dopasuj GLM dla wartości średniej, ale tym razem z wykorzystaniem specyficznego dla każdej obserwacji parametru dyspersji (dzieląc wagę przez parametr dyspersji dla danej obserwacji, uzyskany w poprzednim kroku).
5. Oblicz kwadrat Pearsona lub dewiancję residiów R_i^2 i powtóż poprzednie kroki.

Współdziałanie modeli

Model średniej i model dyspersji są ze sobą ściśle powiązane i wymieniają się informacjami w każdej iteracji:

- **Model średniej dostarcza informacji modelowi dyspersji:** Poprzez obliczenie R_i^2 , model średniej informuje model dyspersji, które obserwacje są bardziej zmienne.
- **Model dyspersji dostarcza informacji modelowi średniej:** Poprzez oszacowane, indywidualne parametry dyspersji (ϕ_i), model dyspersji informuje model średniej, jaką wagę nadać każdej obserwacji. Obserwacjom o wyższej zmienności (większej dyspersji) przypisuje się mniejszą wagę w procesie estymacji modelu dla wartości średniej, co pozwala na ignorowanie szumu i wychwytywanie sygnału.

b) Oznaczanie parametru dyspersji:

Parametr dyspersji φ jest modelowany z użyciem funkcji logarytmicznej jako funkcji T_{mocy} stąd:

$$\hat{\varphi} = \exp \left\{ 0.3762577 - 0.06159272 \cdot \text{nb_RB} + 0.00009304519 \cdot \text{ccm} \right\}$$

Minimalne:

$$\hat{\varphi}_{\min} = \exp \left\{ 0.3762577 - 0.06159272 \cdot 1 + 0.00009304519 \cdot 970 \right\} = 1.500081$$

Maksymalne:

$$\hat{\varphi}_{\max} = \exp \left\{ 0.3762577 - 0.06159272 \cdot 0 + 0.00009304519 \cdot 2700 \right\} = 1.874002$$

Oznaczanie wartości oczekiwanej:

$$\hat{\mu} = \exp \left\{ 6.906079 + 0.00004048551 \cdot \text{ccm} - 0.1907909 \cdot \text{nb_RB} - 0.2025674 \cdot \text{driver_genderMale} \right\}$$

W przypadku z zadania:

$$\hat{\mu} = \exp \left\{ 6.906079 + 0.00004048551 \cdot 2700 - 0.2025674 \right\} = 903.9941$$

Oznaczanie wariancji:

Dla rozkładu gamma, wariancja jest zdefiniowana jako $\hat{\varphi} \cdot \hat{\mu}^2$.

Parametr dyspersji dla tego przypadku:

$$\hat{\varphi} = \exp \left\{ 0.3762577 - 0.06159272 \cdot 0 + 0.00009304519 \cdot 2700 \right\} = 1.874002$$

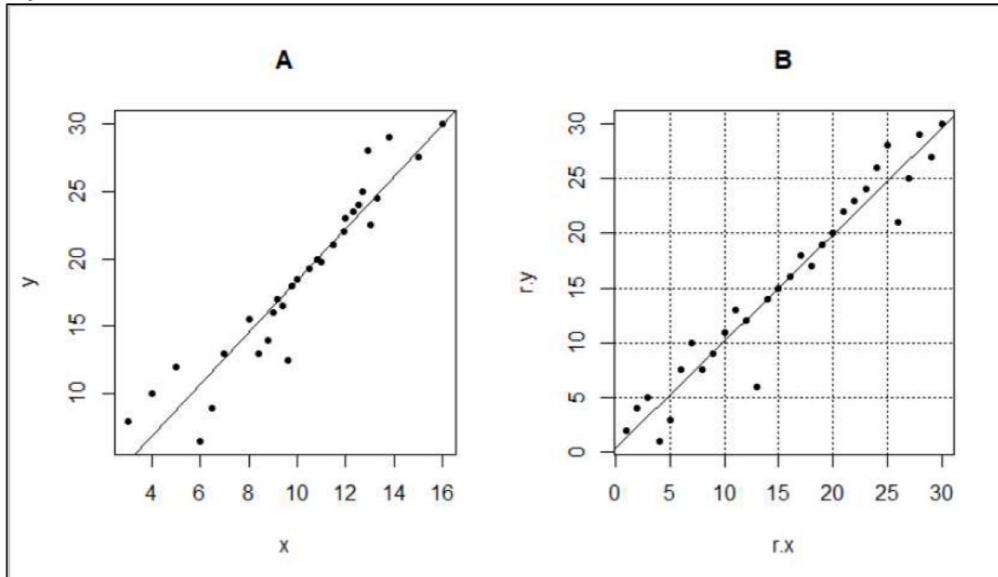
$$\hat{\text{Var}} = 1.874002 \cdot 903.9941^2 = 1531449$$

Zadanie 4.

Zależność między dwiema zmiennymi X i Y analizowano na podstawie 30-sto elementowej próby: $(x_1, y_1) \dots (x_{30}, y_{30})$. Przy czym w próbie tej nie powtarzają się zarówno wartości x_i , jak i wartości y_i . Na rysunku 4.1 przedstawiono wykresy rozrzutu (diagramy korelacyjne) dla danych (wykres A) oraz dla ich rang (wykres B).

- a) (2p.) Na wykresie B liczba niezgodnych par punktów wynosi 25. Oblicz współczynnik korelacji rang Kendalla.

Rys. 4.1



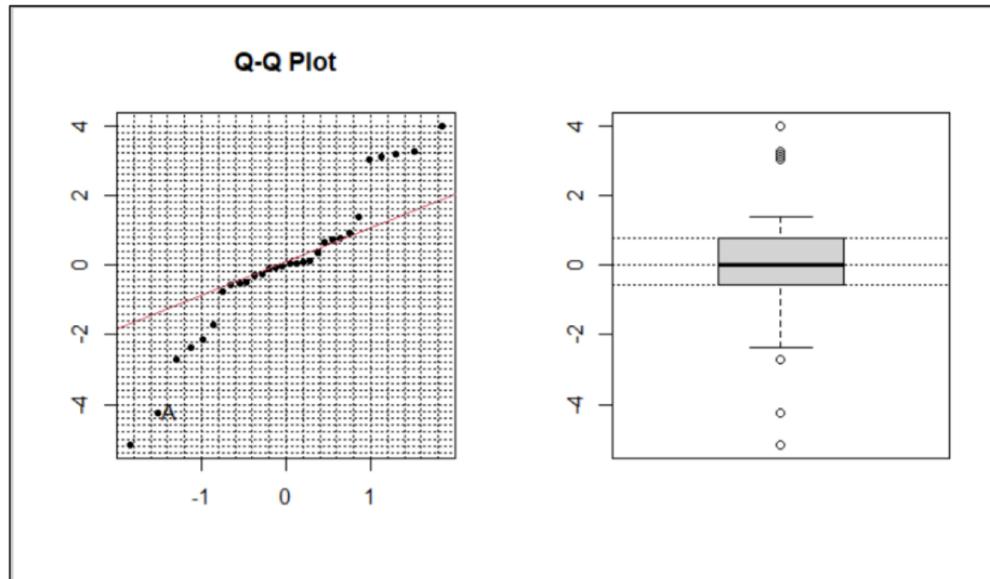
- b) (3p.) Wykorzystując próbę: $(x_1, y_1) \dots (x_{30}, y_{30})$ oszacowano liniowy model regresji zmiennej Y względem X . Otrzymano następujące reszty z tego modelu:

i	1	2	3	4	5	6	7	8	9	10
$\hat{\varepsilon}_i$	3.02	3.1	3.18	-2.71	-4.24	0.33	0.91	-2.36	-2.13	-0.51
i	11	12	13	14	15	16	17	18	19	20
$\hat{\varepsilon}_i$	0.1	-0.78	-5.17	-0.05	0.07	-0.1	0.03	-0.56	-0.32	-0.09
i	21	22	23	24	25	26	27	28	29	30
$\hat{\varepsilon}_i$	0.72	0.64	0.76	1.38	3.99	-1.7	-0.28	3.26	-0.55	0.03

Na rysunku 4.2 przedstawiono wykresy kwantyl-kwantyl (przy założeniu rozkładu normalnego) oraz pudełkowy dla tych reszt.

- Wyjaśnij jakie wartości są na osiach wykresu kwantyl-kwantyl.
- Podaj współrzędne punktu A wskazanego na wykresie kwantyl-kwantyl.
- Na podstawie zamieszczonych wykresów (rys.4.2) wypowiedź się na temat normalności rozkładu reszt. Podaj jeden argument za i jeden przeciw temu założeniu.

Rys. 4.2



$$a) \hat{t} = 1 - \frac{\frac{2Q}{n(n-1)}}{\frac{2}{2}} \quad n - liczebność próby$$

Q - limba par niezgodnych

$$\hat{t} = 1 - \frac{\frac{2 \cdot 25}{30 \cdot 29}}{\frac{2}{2}} = 0.885$$

b) Dla X: kwantyl standaryzowanego rozkładu normalnego $\mu_{\frac{2}{31}}$, gdzie
 $i = 1, \dots, 30$

Dla Y: upomiedlowane resty

Współmiediane punktu A:

$$X = \mu_{\frac{2}{31}} \quad 1 - \frac{2}{31} = 0.935 \quad \text{kompatując z tabeli } \mu_{\frac{2}{31}} = -1.52$$

$y = -4.54$ dając wartości upomiedlowanych rest

$$A = (-1.52, -4.54)$$

Preciw: obserwacje odstające na wykresie produktywnym, punkty oddzielające na wykresie QQ w sugerują gąbkę ogony.

za: w rejonie centralnej wykresu QQ punkty nie odbiegają od wartości oznaczonej.

Zadanie 5.

- (1p.) Podaj definicję procesu GARCH(p, q).
- (2p.) Do jakich celów służą modele klasy GARCH?
- (2p.) Na podstawie szeregu czasowego liczącego 1837 obserwacji stóp zwrotu (r_t) PZU (od 2016-06-06 do 2023-10-05) oszacowano model ARMA(0,0)-GARCH(1,1) z gaussowskimi innowacjami. Uzyskano następujące wyniki:

Optimal Parameters

	Estimate	Std. Error	t value	Pr(> t)
mu	0.000366	0.000386	0.94631	0.343988
omega	0.000021	0.000006	3.65831	0.000254
alpha1	0.077897	0.014377	5.41818	0.000000
beta1	0.854787	0.028546	29.94420	0.000000

Stopy zwrotu r_t i oszacowane warunkowe odchylenia standardowe $\hat{\sigma}_t$ dla 3 ostatnich obserwacji przedstawia tabela 5.1.

Tab. 5.1

t	1835	1836	1837
r_t	0.00497761	0.00692729	-0.00049322
$\hat{\sigma}_t$	0.01856621	0.01781931	0.01720450

Na podstawie powyższych informacji oszacuj na okres $t = 1838$:

- warunkowe odchylenie standardowe,
- jednodniowy VaR (wartość zagrożoną), przyjmując poziom tolerancji $\alpha = 0.05$.

a)

Definicja procesu GARCH(p, q)

Niech $(Z_t)_{t \in \mathbb{Z}}$ będzie procesem typu **ścisły biały szum SWN(0, 1)**. Proces $(X_t)_{t \in \mathbb{Z}}$ jest procesem GARCH(p, q), jeśli jest ścisłe stacjonarny i dla każdego $t \in \mathbb{Z}$ oraz dla pewnego procesu $(\sigma_t)_{t \in \mathbb{Z}}$ o wartościach dodatnich spełnia równania:

$$X_t = \sigma_t Z_t$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^p \alpha_i X_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2$$

gdzie parametry spełniają warunki: $\alpha_0 > 0$, $\alpha_i \geq 0$ dla $i = 1, \dots, p$ oraz $\beta_j \geq 0$ dla $j = 1, \dots, q$. W tym modelu σ_t^2 jest warunkową wariancją procesu X_t .

b)

Zastosowanie modeli GARCH

Modele klasy GARCH służą przede wszystkim do **modelowania i prognozowania zmienności** (ang. *volatility*) finansowych szeregów czasowych. Ich głównym celem jest uchwycenie kluczowych empirycznych właściwości danych finansowych, zwanych stylizowanymi faktami, których nie uwzględniają prostsze modele.

- **Modelowanie klasteryzacji zmienności:** Jest to tendencja obserwowana w danych finansowych, gdzie okresy dużych wahań cen następują po sobie, a okresy spokoju również występują w seriach. Modele GARCH odwzorowują to zjawisko, pozwalając, aby dzisiejsza wariancja warunkowa (σ_t^2) zależała od wcześniejszych kwadratów obserwacji (X_{t-1}^2) oraz od wcześniejszej wariancji warunkowej (σ_{t-1}^2).
- **Uchwycenie grubych ogonów (leptokurtozy):** Nawet jeśli innowacje (Z_t) w procesie GARCH pochodzą z rozkładu normalnego, stacjonarny rozkład samego procesu (X_t) jest leptokurtyczny (ma "grubsze ogony" niż rozkład normalny), co jest zgodne z empirycznymi właściwościami stóp zwrotu z aktywów finansowych.
- **Prognozowanie zmienności i szacowanie miar ryzyka:** Dopasowane modele GARCH pozwalają na prognozowanie przyszłej zmienności warunkowej. Prognozy te są kluczowym elementem w szacowaniu miar ryzyka, takich jak **Value-at-Risk (VaR)** i **Expected Shortfall (ES)**, w ujęciu warunkowym, czyli uwzględniającym najnowsze informacje rynkowe.

c) Przyjmując oznaczenia z zadania :

$$\hat{\sigma}_t^2 = \omega + \sum_{i=1}^p \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^q \beta_j \hat{\sigma}_{t-j}^2$$

$$p = q = 1$$

$$\hat{\sigma}_{1837}^2 = \omega + \alpha_1 \epsilon_{1837}^2 + \beta_1 \hat{\sigma}_{1837}^2$$

ϵ_{1837} - różnica między mierzoną stopą zwrotu a jej wartością oczekiwanej

$$\epsilon_{1837} = r_{1837} - \mu$$

$$\hat{\sigma}_{1837}^2 = \omega + \alpha (\epsilon_{1837} - \mu) + \beta \hat{\sigma}_{1837}^2$$

$$\begin{aligned} \hat{\sigma}_{1837}^2 &= 0.000021 + 0.077897 (-0.00049322 - 0.000366)^2 + \\ &+ 0.854787 \cdot 0.01720450 = 0.0002742922 \end{aligned}$$

$$\hat{\sigma}_{1837} = \sqrt{0.0002742922} = 0.01656177$$

Oznaczenia VaR :

$$VaR_\alpha = \mu + \hat{\sigma}_{t+1} \cdot z_\alpha \quad z_\alpha - kwantyl z rozkładu normalnego$$

$$VaR_{0.05} = 0.000366 + 0.01656177 \cdot (-1.64) = -0.0267953$$

Zadanie 6.

Wykorzystując dane zwarte w tabeli 6.1, gdzie symbolem (*) oznaczono obserwacje cenzurowane z góry:

- (3p.) Skonstruuj estymator Nelsona-Åalena dla funkcji przeżycia $S(x)$. Uwzględnij poprawkę Kleina-Moeschbergera, przyjmując $\gamma = 22$.
- (2p.) Oszacuj wariancję estymatora Nelsona-Åalena dla $S(2)$.

Tab. 6.1

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
x_i	1	2	3*	4	4	4*	4*	5	7*	8	8	8	9	9	9	9	10*	12	12	15*

a) Estymator Nelsona - Åalena :

$$\hat{S}_n(y) = \exp \left\{ -\hat{H}_n(y) \right\}, \text{ gdzie } \hat{H}_n(y) = \sum_{i: x_i \leq y} \frac{1}{n_i}$$

i	x_i	n_i	b_i	n_i	$\hat{H}_n(x_i)$	$\hat{S}_n(x_i) = \exp \left\{ -\hat{H}_n(x_i) \right\}$
1	1	1	0	20	$\frac{1}{20} = 0.05$	0.951
2	2	1	1	19	$0.05 + \frac{1}{19} = 0.1026$	0.902
3	4	2	2	17	$0.1026 + \frac{2}{17} = 0.2202$	0.802
4	5	1	1	13	$0.2202 + \frac{1}{13} = 0.2971$	0.743
5	8	3	0	11	$0.2971 + \frac{3}{11} = 0.57$	0.566
6	9	4	1	8	$0.57 + \frac{4}{8} = 1.07$	0.343
7	12	2	1	3	$1.07 + \frac{2}{3} = 1.7367$	0.176

Poprawka Kleina - Moeschbergera, dla $x \geq x_{max}$, gdzie $x_{max} = 15$ to najniższa obserwacja cenzurowana, funkcja przeżycia jest ekstrapolowana.

Dla $\gamma = 22$:

$$\hat{S}_n(x) = \hat{S}_n(12) = 0.176 \text{ dla } 15 \leq y < 22$$

$$\hat{S}_n(x) = 0 \text{ dla } y \geq 22$$

$$\hat{F}_n(x) = \begin{cases} 1 & , x < 1 \\ 0.951 & , 1 \leq x < 2 \\ 0.902 & , 2 \leq x < 4 \\ 0.802 & , 4 \leq x < 5 \\ 0.743 & , 5 \leq x < 8 \\ 0.566 & , 8 \leq x < 9 \\ 0.343 & , 9 \leq x < 12 \\ 0.176 & , 12 \leq x < 22 \\ 0 & , x \geq 22 \end{cases}$$

b) Jneba shomystar z estymetora Mleima:

$$\hat{Var}[\hat{F}_n(y)] = [\hat{F}_n(y)]^2 \sum_{i:y_i \leq y} \frac{i(n_i - 1)}{n_i^3}, \quad y < y_{\max}$$

$$\hat{Var}[\hat{F}_{20}(2)] = 0.902^2 \left[\frac{1 \cdot 19}{20^3} + \frac{1 \cdot 18}{19^3} \right] = 0.00407$$

Zadanie 7.

- a) (1p.) Krótko przedstaw ideę statystycznych metod uczenia zespołowego (*Ensemble Statistical Learning*).
- b) (1p.) Wymień co najmniej trzy takie metody wykorzystujące drzewa (*Tree Ensemble Methods*)
- c) (3p.) Opisz jedną metodę spośród wymienionych w punkcie b).

a)

Idea statystycznych metod uczenia zespołowego (*Ensemble Statistical Learning*)

Metody uczenia zespołowego to podejścia, które **łączą wiele prostych modeli, zwanych "słabyimi uczniami" (weak learners), w celu uzyskania jednego, potężnego modelu predykcyjnego**. Zamiast polegać na jednym, skomplikowanym modelu, metody te agregują predykcje z wielu prostszych modeli. Głównym celem jest **zmnieszenie wariancji** i poprawa dokładności predykcyjnej w porównaniu do pojedynczego modelu. Poprzez uśrednienie wyników z wielu modeli zbudowanych na różnych próbkach danych, ogólny wynik staje się bardziej stabilny i mniej podatny na specyfikę pojedynczego zbioru treningowego.

b)

Metody zespołowe wykorzystujące drzewa (*Tree Ensemble Methods*)

- **Bagging** (agregacja bootstrapowa).
- **Lasy losowe** (Random Forests).
- **Boosting**.
- **Bayesowskie addytywne drzewa regresyjne** (Bayesian Additive Regression Trees, BART).

c)

Opis Bagging

Bagging, czyli agregacja bootstrapowa (ang. *bootstrap aggregation*), to ogólna procedura mająca na celu zmniejszenie wariancji w metodach uczenia statystycznego, szczególnie skuteczna w przypadku drzew decyzyjnych, które charakteryzują się wysoką wariancją.

Mechanizm działania opiera się na idei, że uśrednianie zbioru obserwacji redukuje wariancję. Ponieważ zazwyczaj nie mamy dostępu do wielu niezależnych zbiorów treningowych, Bagging tworzy je sztucznie za pomocą techniki **bootstrapu**, czyli losowania ze zwracaniem z oryginalnego zbioru danych.

Proces składa się z następujących kroków:

1. **Tworzenie zbiorów bootstrapowych:** Z oryginalnego zbioru treningowego o n obserwacjach tworzy się B nowych zbiorów treningowych, każdy o rozmiarze n , poprzez losowanie ze zwracaniem. Każdy z tych zbiorów jest nieco inny.
2. **Budowanie modeli:** Na każdym z B "bootstrapowych" zbiorów danych budowane jest osobne, **głębokie i nieprycinane drzewo decyzyjne**. Każde z tych drzew ma niską obciążalność (bias), ale wysoką wariancję.

3. Agregacja predykcji: Aby uzyskać ostateczną predykcję dla nowej obserwacji, wyniki z **B** drzew są łączone:

- W przypadku **regresji** (odpowiedź ilościowa), predykcje są uśredniane.
- W przypadku **klasyfikacji** (odpowiedź jakościowa), ostateczna predykcja jest wynikiem **głosowania większościowego** – wybierana jest klasa najczęściej wskazywana przez poszczególne drzewa.

Zalety i wady

Główną zaletą metody Bagging jest **znacząca poprawa dokładności predykcyjnej** w porównaniu do pojedynczego drzewa, dzięki redukcji wariancji. Użycie dużej liczby drzew **B** nie prowadzi do przeuczenia (overfittingu).

Podstawową wadą jest **utrata interpretoalności**. Zamiast jednego, łatwego do zwizualizowania drzewa, otrzymujemy setki lub tysiące drzew, których nie da się przedstawić w prostej, graficznej formie.

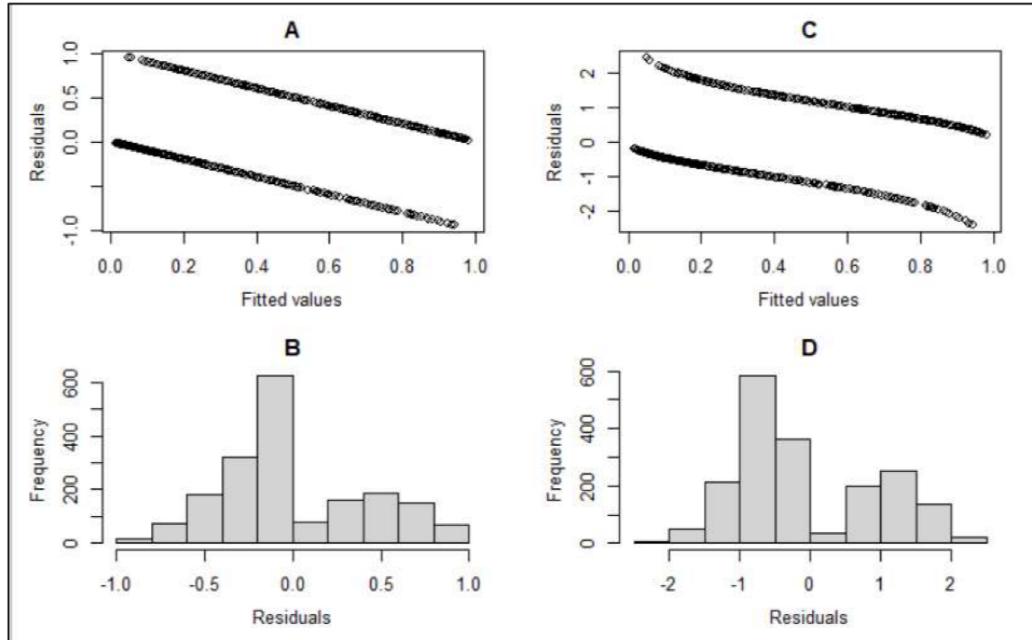
Estymacja błędu Out-of-Bag (OOB)

Bagging oferuje efektywny sposób estymacji błędu testowego bez konieczności stosowania walidacji krzyżowej. Każde drzewo jest budowane na około 2/3 unikalnych obserwacji z oryginalnego zbioru. Pozostała 1/3 obserwacji, niewykorzystana do budowy danego drzewa, to tzw. obserwacje **Out-of-Bag (OOB)**. Dla każdej obserwacji w zbiorze treningowym można uzyskać predykcję OOB, uśredniając wyniki tylko z tych drzew, dla których była ona obserwacją OOB. Błąd OOB, obliczony na podstawie tych predykcji, jest wiarygodnym estymatorem błędu testowego.

Zadanie 8.

- (1p.) W jaki sposób oblicza się reszty dewiancyjne (*deviance residuals*)?
- (2p.) Dlaczego są one często preferowaną formą reszt w modelach GLM? Jakie są główne zalety ich wykorzystania w porównaniu z innymi rodzajami reszt?
- (2p.) Na rysunku 8.1 przedstawiono wykresy reszt zwykłych i dewiancyjnych w zależności od wartości dopasowanych oraz ich histogramy. Reszty te odpowiadają temu samemu oszacowanemu uogólnionemu modelowi liniowemu. Wskaż i uzasadnij, jaki rozkład miała zmienne objaśniana w tym modelu oraz któremu rodzajowi reszt odpowiadają poszczególne rysunki A, B, C i D.

Rys. 8.1



a) $d(y, \hat{\theta}) = 2(y \Theta_y - \alpha(\Theta_y) - (y \hat{\theta} - \alpha(\hat{\theta})))$ – kontynuacja odpowiedzi y do dewiancji, Θ_y jest wartością parametru hanomicznego spełniającego zależność $\alpha'(\Theta_y) = y$

Dewiancja:

$$D(y, \hat{\mu}) = \sum_{i=1}^n d_i, \text{ gdzie } d_i = d(y_i, \hat{\theta}_i)$$

Reszty dewiancyjne:

$$r_i^D = \text{sign}(y_i - \hat{\mu}_i) \sqrt{d_i}, \text{ gdzie } \text{sign}(y_i - \hat{\mu}_i) = \begin{cases} 1 & \text{jeli } y_i > \hat{\mu}_i \\ -1 & \text{p.p.} \end{cases}$$

Reszty dewiancyjne oblicza się jako pierwiastki z właściwych danych obserwacji do tejnej dewiancji modelu, z uwzględnieniem zmiany reszty prostej. y_i – wartość obserwowana, $\hat{\mu}_i$ – wartość dopasowana przez model.

b)

Reszty dewiancyjne są często preferowaną formą w Uogólnionych Modelach Liniowych (GLM), ponieważ w przeciwieństwie do innych typów reszt, uwzględniają one kształt rozkładu zmiennej odpowiedzi, w tym jego skośność. Jest to kluczowe w zastosowaniach aktuarialnych, gdzie rozkłady rzadko są symetryczne.

Główne zalety ich wykorzystania w porównaniu z innymi rodzajami reszt to:

- **Lepsze dopasowanie do założeń GLM:** W przeciwieństwie do reszt Pearsona, które wywodzą się z modeli liniowych, reszty dewiancyjne są bezpośrednio powiązane z funkcją wiarygodności przyjętego rozkładu z rodziną wykładniczej, co czyni je bardziej odpowiednimi dla struktury GLM.
- **Wykrywanie obserwacji słabo dopasowanych:** Pozwalają zidentyfikować te obserwacje, które w największym stopniu przyczyniają się do wzrostu dewiancji, a więc wskazują na miejsca, w których model niedostatecznie dobrze dopasowuje się do danych.
- **Korekta na heteroskedastyczność:** Podobnie jak reszty Pearsona, ale w przeciwieństwie do reszt prostych (surowych), korygują one fakt, że wariancja w modelach GLM często zależy od wartości średniej. Reszty proste mogą być przez to mniej informatywne.
- **Lepsza interpretacja dla danych dyskretnych:** Chociaż indywidualne reszty dla danych dyskretnych (np. liczby szkód) mogą być trudne do interpretacji, reszty dewiancyjne obliczone dla zagregowanych grup ryzyka dają znacznie lepsze wskazówki co do poprawności dopasowania modelu.

c) A i B - zwykłe reszty

C i D - reszty dewiancyjne

W rozważanym przypadku, zwykłe reszty muszą mieścić się w przedziale $(-1, 1)$, tq to różnice między zaobserwowanymi wartościami zmiennej zależnej tj. 0 lub 1, a oszacowanymi prawdopodobieństwami.

Zadanie 9.

- a) (3p.) Opisz koncepcję redukcji wariancji w metodzie Monte Carlo za pomocą metody próbkowania ważonego (metody *IS - importance sampling*).
- b) (2p.) Zmienna losowa X ma standardowy rozkład normalny ($X \sim N(0, 1)$). Wykorzystując przesunięty rozkład wykładniczy z parametrem $\lambda = 1$ i odpowiednio dobranym parametrem przesunięcia, wyznacz estymator Monte Carlo wykorzystujący próbkowanie ważne (*importance sampling estimator*) dla prawdopodobieństwa $P(X > 3.5)$.

Q)

Próbkowanie ważne (Importance Sampling - IS) to technika redukcji wariancji w metodzie Monte Carlo, która polega na zastąpieniu oryginalnego rozkładu prawdopodobieństwa innym, który koncentruje większe prawdopodobieństwo w regionach o największym znaczeniu dla estymowanej wartości. Celem jest zwiększenie wydajności symulacji, zwłaszcza przy szacowaniu prawdopodobieństw zdarzeń rzadkich lub wartości w ogonach rozkładu.

Kluczowe założenia

Podstawowym problemem w prostej metodzie Monte Carlo jest jej niska wydajność, gdy wiele generowanych prób losowych nie trafia w interesujący nas obszar. Próbkowanie ważne rozwiązuje ten problem poprzez zmianę gęstości prawdopodobieństwa, z której losowane są próby, z oryginalnej $f_X(x)$ na nową gęstość $\tilde{f}_X(x)$.

Aby wynik pozostał nieobciążony, każda wylosowana próba musi zostać skorygowana przez pomnożenie jej przez **iloraz wiarygodności (likelihood ratio)**, dany wzorem $f_X(x)/\tilde{f}_X(x)$.

Estymator i redukcja wariancji

Wartość oczekiwana $E(X)$ można przedstawić jako:

$$E(X) = \int x f_X(x) dx = \int \left(x \frac{f_X(x)}{\tilde{f}_X(x)} \right) \tilde{f}_X(x) dx = E \left(\tilde{X} \frac{f_X(\tilde{X})}{\tilde{f}_X(\tilde{X})} \right)$$

gdzie \tilde{X} jest zmienną losową o gęstości $\tilde{f}_X(x)$.

Estymator próbkowania ważonego dla n symulacji ma postać:

$$\hat{\mu}_{In} = \frac{1}{n} \sum_{i=1}^n \tilde{X}_i \frac{f_X(\tilde{X}_i)}{\tilde{f}_X(\tilde{X}_i)}$$

Estymator ten jest nieobciążony, a jego wariancja jest mniejsza niż wariancja standardowego estymatora Monte Carlo, jeśli nowa gęstość $\tilde{f}_X(x)$ jest dobrana tak, aby iloraz wiarygodności był mały tam, gdzie wartość $x^2 f_X(x)$ jest duża.

b) Estymator :

$$\hat{P}(A) = \frac{1}{m} \sum_{i=1}^m \frac{f(X_i)}{g(X_i)} h(X_i), \text{ gdzie próba } X_1, X_2, \dots, X_m \text{ jest losowana z rozkładu } g.$$

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}x^2\right] \quad - \text{ normal } N(0, 1)$$

$$g(x) = \exp\left[-(x-3.5)\right] \quad - \text{ prawdopodobieństwo wykroczenia } \geq \lambda = 1$$

$$h(x) = \mathbb{1}_{x>3.5} = 1 \quad - \text{ funkcja równe jedna 1 bo losujemy } \geq \text{ wartości prawdopodobego o } 3.5$$

Po ustaleniu do wron:

$$\hat{P}(X > 3.5) = \frac{1}{n} \sum_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}x_i^2 + x_i - 3.5\right]$$

Zadanie 10.

- a) (3p.) Przedstaw przebieg procesu grupowania hierarchicznego według algorytmu aglomeracyjnego.
- b) (2p.) Analizowano podobieństwo profili ryzyka czterech portfeli ubezpieczeń: A, B, C i D z wykorzystaniem aglomeracyjnego grupowania hierarchicznego, w którym odległość między skupieniami (klastrami) mierzono metodą pełnego wiązania (*complete linkage*). Każdy portfel opisano pięcioma zmiennymi diagnostycznymi. W oparciu o zestandaryzowane wartości tych zmiennych otrzymano następującą macierz niepodobieństwa:

	A	B	C	D
A		0.3	0.4	0.7
B	0.3		0.5	0.8
C	0.4	0.5		0.45
D	0.7	0.8	0.45	

Narysuj dendrogram, będący wynikiem tego grupowania. Zaznacz na nim wysokość węzłów i podpisz liście.

a)

Grupowanie hierarchiczne to metoda analizy skupień, która buduje hierarchię klastrów. Podejście aglomeracyjne, zwane również podejściem "oddolnym" (bottom-up), jest najpowszechniejszą metodą grupowania hierarchicznego.

Proces ten tworzy wizualną reprezentację danych w postaci drzewa zwaną **dendrogramem**, który ilustruje, jak obserwacje są kolejno łączone w klastry.

Algorytm grupowania aglomeracyjnego

Algorytm przebiega iteracyjnie, łącząc obserwacje i klastry krok po kroku, od pojedynczych punktów aż do jednego, dużego klastra obejmującego cały zbiór danych.

Krok 1: Inicjalizacja

Proces rozpoczyna się od zdefiniowania miary odmiенноści (najczęściej odległości euklidesowej) między każdą parą obserwacji w zbiorze danych. Na tym etapie każda z **n** obserwacji jest traktowana jako osobny, jednoelementowy klaster.

Krok 2: Iteracyjne łączenie klastrów

Algorytm wykonuje **n-1** kroków, aby połączyć wszystkie obserwacje w jeden klaster. W każdej iteracji:

- 1. Identyfikacja najbliższych klastrów:** Algorytm identyfikuje dwa klastry, które są do siebie najbardziej podobne (najmniej odmienne).
- 2. Fuzja (łączenie) klastrów:** Te dwa najbliższe klastry są łączone w jeden nowy klaster.
- 3. Aktualizacja miar odmiенноści:** Obliczane są nowe miary odmiенноści między nowo powstałym klastrem a wszystkimi pozostałymi.

Proces ten jest kontynuowany, aż wszystkie obserwacje zostaną połączone w jeden klaster.

Wynik: Dendrogram

Wynikiem działania algorytmu jest **dendrogram** – struktura przypominająca drzewo, która wizualizuje proces łączenia.

- Liście** na dole dendrogramu reprezentują poszczególne obserwacje.
- Wysokość**, na której dwa klastry są łączone, reprezentuje stopień ich odmiенноści. Im niżzej następuje połączenie, tym bardziej podobne są obserwacje.
- Aby uzyskać określoną liczbę klastrów, można "przeciąć" dendrogram na odpowiedniej wysokości. Liczba linii dendrogramu przeciętych przez tę poziomą linię cięcia odpowiada liczbie uzyskanych klastrów.

b) 1. Wyliczamy najmniejsze odległość w skupinie. Jest to 0.3 dla pary (A, B), teraz jest to nasze piąte skupienie na poziomie 0.3.

2. Liczymy odległość (A, B) od pozostałych portfeli (complete linkage czyli odległość między najdalejymi elementami skupisk)

$$d\{(A, B), C\} = \max\{d(A, C), d(B, C)\} = \max(0.4, 0.5) = 0.5$$

$$d\{(A, B), D\} = \max\{d(A, D), d(B, D)\} = \max(0.7, 0.8) = 0.8$$

Jedno z pięciu obliczonych odległości między pozostałymi portfelami czyli:

$$d\{C, D\} = 0.45$$

3. Wybieramy najmniejszą odległość i tworzymy elementy w skupinie:

$$d\{C, D\} = 0.45 \text{ myli nowe skupienie to } (C, D).$$

4. Znowu liczymy odległości między nowymi skupinami:

$$\begin{aligned} d\{(A, B), (C, D)\} &= \max\{d(A, C), d(A, D), d(B, C), d(B, D)\} = \\ &= \max\{0.4, 0.4, 0.5, 0.8\} = 0.8 \end{aligned}$$

Liczymy oba skupienia na wysokości 0.8, kończąc proces.

