

- a) (2p.) Co to jest dewiancja (*deviance*) w kontekście uogólnionych modeli liniowych (GLM)?
- b) (1p.) Jak wygląda ogólna postać wzoru na dewiancję i co oznaczają jego poszczególne składniki?
- c) (2p.) Na podstawie danych z pierwszych dwóch kolumn tabeli 1.1 oszacowano uogólniony model liniowy, zakładając, że zmienna objaśniana Y podlega rozkładowi Poissona oraz stosując link kanoniczny. Uzyskane wyniki przedstawiono poniżej:

```
call:  
glm(formula = Y ~ 1 + X, family = poisson)  
  
Coefficients:  
            Estimate Std. Error z value  
(Intercept) 1.79176   0.23570  7.602  
X            0.09531   0.29301  0.325  
---
```

Tab. 1.1

$x_i$	$y_i$	$\hat{y}_i$	$y_i \ln\left(\frac{y_i}{\hat{y}_i}\right)$
(kolumna 1)	(kolumna 2)	(kolumna 3)	(kolumna 4)
0	7	?	?
0	9	6.0	3.6492
0	2	6.0	-2.1972
1	3	6.6	-2.3654
1	10	6.6	4.1552
1	8	6.6	1.5390
1	5	6.6	-1.3882
1	7	6.6	0.4119

Wykorzystując informacje podane w tabeli 1.1 oblicz dewiancję tego modelu.

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 rodziną GLM. Kwantyfikuje ona zmienność w danych, która nie została wyjaśniona przez analizowany model.

Koncepcja dewiancji opiera się na porównaniu dopasowania dwóch modeli:

- Modelu analizowanego: Modelu, którego jakość dopasowania chcemy ocenić.
- Modelu nasyconego (pełnego): Modelu, który idealnie pasuje do danych, posiadając tyle samo parametrów, co obserwacji. W tym modelu wartość dopasowana dla każdej obserwacji jest równa jej wartości obserwowanej ( $\hat{\mu} = y_i$ ). Jest to najlepsze możliwe dopasowanie, jakie można uzyskać dla danej rodziny rozkładów.

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

$y$  - dane obserwowane

$\hat{\mu}$  - wielotor wartości prawidłowych po model

$\varphi$  - parametr dyspersji, jest to stała, która studiuje wariancję rozkładu

$L_{full}$  - logarytm funkcji prawdopodobieństwa dla modelu nasyconego

$L(\hat{\beta})$  - logarytm funkcji prawdopodobieństwa dla dopasowanego modelu

$\hat{\beta}$  - parametry występujące w naszym modelu

c) Wartość prawidłowa w kolumnie 3 dla modelu Poissona może być obliczona ze wzoru:

$$\hat{y}_i := \exp[\beta_0 + \beta_1 \cdot x_i]$$

$$\beta_0 = 1.49176 \quad \beta_1 = 0.09531 \quad x_1 = 0$$

$$\hat{y}_1 = \exp[1.49176] \approx 6.0$$

Wartości w kolumnie 4:

$$y_i \ln\left(\frac{y_i}{\hat{y}_i}\right) = 7 \cdot \ln\left(\frac{7}{6}\right) \approx 1.0791$$

Wpływ na deniancje w modelu Poissona:

$$D(y, \hat{y}) = 2 \sum_{i=1}^m \left[ y_i \ln\left(\frac{y_i}{\hat{y}_i}\right) - (y_i - \hat{y}_i) \right] = 2 \sum_{i=1}^m \left[ y_i \ln\left(\frac{y_i}{\hat{y}_i}\right) \right] - \sum_{i=1}^m (y_i - \hat{y}_i)$$

W modelu Poissona drugi człon w powyższym równaniu zawsze wynosi 0.

$$D(y, \hat{y}) = 2 \sum_{i=1}^m \left[ y_i \ln \left( \frac{y_i}{\hat{y}_i} \right) \right] = 2 \cdot 4.836 = 9.672$$

**Zadanie 2.**

- a) (2p.) Na czym polega metoda Hilla w estymacji ogona rozkładu?
- b) (1p.) W jaki sposób estymator Hilla może być wykorzystany do modelowania dużych szkód w ubezpieczeniach majątkowych?
- c) (2p.) Aby oszacować ogon rozkładu wysokości szkód  $X$  (w mln zł), aktuarusz stosuje metodę Hilla. Tabela 2.1 przedstawia uporządkowane wartości szkód  $X_{k,n}$  (od największej do najmniejszej) oraz oszacowania indeksu ogona  $\hat{\alpha}_{k,n}$ , natomiast odpowiedni wykres Hilla został zaprezentowany na rysunku 2.1. Korzystając z estymatora ogona Hilla (*Hill tail estimator*), oszacuj prawdopodobieństwo, że szkody przekroczą 4 mln zł.

Tab. 2.1

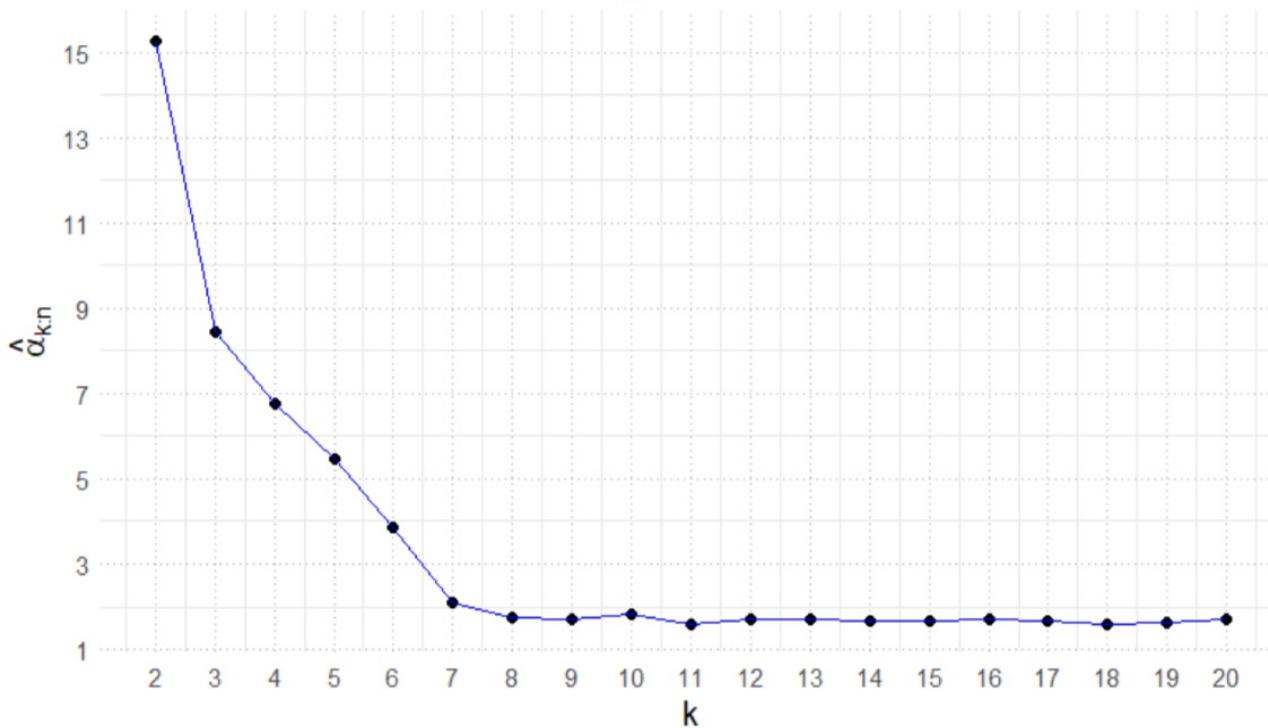
$k$	$X_{k,n}$	$\hat{\alpha}_{k,n}$
1	4.922	
2	4.318	15.258
3	3.858	8.426
4	3.567	6.757
5	3.292	5.483
6	2.899	3.876
7	2.142	2.081
8	1.804	1.752
9	1.653	1.709
10	1.606	1.810
11	1.400	1.594
12	1.380	1.700
13	1.319	1.711
14	1.247	1.682
15	1.187	1.663
16	1.153	1.691
17	1.102	1.671
18	1.032	1.594
19	1.018	1.645
20	1.004	1.695

Uwaga! Estymator ogona Hilla (*Hill tail estimator*) ma postać:

$$\hat{F}(x) = \frac{k}{n} \left( \frac{x}{X_{k,n}} \right)^{-\hat{\alpha}_{k,n}}, x \geq X_{k,n}$$

Rys. 2.1

Wykres Hilla



Q)

Metoda Hilla jest alternatywą dla modelowania opartego na uogólnionym rozkładzie Pareto (GPD), przeznaczoną specjalnie dla rozkładów ciężkoogonowych (należących do dziedziny przyciągania rozkładu Frécheta).

Główne punkty:

- Cel metody: głównym celem Metody Hilla nie jest modelowanie całej dystrybucji nadwyżek ponad próg, lecz bezpośrednie oszacowanie indeksu ogona ( $\alpha$ ). Indeks ten opisuje, jak szybko maleje prawdopodobieństwo w ogonie rozkładu.
- Estymator Hilla: estymacja odbywa się za pomocą estymatora Hilla, który jest funkcją  $k$  największych statystyk pozycyjnych z próby  $n$  obserwacji.
- Wykres Hilla: w praktyce, kluczowym narzędziem jest wykres Hilla, który przedstawia wartość estymowanego indeksu ogona w zależności od liczby użytych statystyk pozycyjnych ( $k$ ). Analityk poszukuje na wykresie stabilnego regionu, który sugeruje właściwą wartość indeksu ogona. Interpretacja wykresu może być jednak trudna.
- Estymacja ogona rozkładu: wyestymowany indeks ogona jest następnie używany do skonstruowania estymatora ogona Hilla, który pozwala szacować prawdopodobieństwa w ogonie rozkładu strat powyżej progu.

### Odp. b)

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.

c) Należy wybrać próg  $k$  który od której obserwacji mierzymy ogon rozkładu.

W tym celu analizujemy wykres Hilla, sprawiamy na nim region, w którym wartość estymatora  $\hat{\alpha}_{n,m}$  przestaje gwałtownie spadać.

Wykres wypłaszcza się od  $k=8$ .

Odczytujemy dane z tabeli:

$$\lambda_{e,n} = 1.752$$

$$X_{e,n} = 1.804$$

$$\hat{F}(4) = \frac{4}{20} \left( \frac{4}{1.804} \right)^{-1.752} = 0.0991$$

### Zadanie 3.

Dla dziewięciu osób, obserwowanych od urodzenia, dostępne są następujące dane dotyczące przeżycia:

$$27, 30^*, 34, 58^*, 68, 68^*, 70, 77, 78^*.$$

Symbolom (\*) oznaczono obserwacje cenzurowane z góry.

- a) (3p.) Wykorzystując metodę Kaplana-Meiera oblicz  $\hat{S}_9(68)$  oraz oszacuj odchylenie standardowe dla  $\hat{S}_9(68)$ .
- b) (2p.) Wykorzystując metodę Nelsona-Aalena oblicz  $\hat{S}_9(68)$ .

a) Estymator Kaplana - Meiera :

$$\hat{I}_n(y) = \prod_{i:y_i \leq y} \left(1 - \frac{s_i}{n_i}\right)$$

$i$	$y_i$	$s_i$	$b_i$	$n_i$	$\hat{I}_n(y_i)$
1	27	1	1	9	$1 - \frac{1}{9} = 0.8889$
2	34	1	1	7	$0.8889 \cdot (1 - \frac{1}{7}) = 0.7619$
3	68	1	1	5	$0.7619 \cdot (1 - \frac{1}{5}) = 0.6095$

$y_i$  - czas, w którym nastąpiło  $i$ -te zdarzenie

$s_i$  - liczba zgonów w wieku  $y_i$

$b_i$  - liczba zgonów cenzurowanych po  $y_i$  a przed  $y_{i+1}$

$n_i$  - liczba osób w grupie ryzyka tui przed  $y_i$

Pred  $y_1 = 27$  jest  $n_1 = 9$  osób, następuje  $s_1 = 1$  zgon oraz  $b_1 = 1$

zgony cenzurowane miedzy  $y_1 = 27$  a  $y_2 = 34$ ,  $n_2 = n_1 - s_1 - b_1$  itd.

Do obliczenia wariancji miarymowa Greenwoda :

$$\text{Var}(\hat{I}_n(y)) = \left[ \hat{I}_n(y) \right]^2 \sum_{\substack{i \\ y_i \leq y}} \frac{s_i}{n_i(n_i - s_i)}$$

$$\text{Var}(\hat{I}_9(68)) = 0.6095^2 \cdot \left[ \frac{1}{9(9-1)} + \frac{1}{7(7-1)} + \frac{1}{5(5-1)} \right] = 0,03258$$

$$\text{SD}(\hat{f}_g(68)) = \sqrt{0,03258} = 0,1805$$

b) Estimator Nelson - Tahara:

$$\hat{f}_n(y) = \exp\{-\hat{H}_n(y)\}, \text{ gdzie } \hat{H}_n(y) = \sum_{i:y_i \leq y} \frac{1}{n_i}$$

Konystejec z poprzednich tabellii:

$$\hat{H}_g(68) = \frac{1}{9} + \frac{1}{7} + \frac{1}{5} = 0.4540$$

$$\hat{f}_g(68) = \exp\{-0.4540\} = 0.6351$$

#### Zadanie 4.

W ciągu roku odnotowano 20 następujących szkód  $x_i$  (w mln EUR) związanych z ekstremalnymi zjawiskami pogodowymi:

1, 1, 1, 1, 1, 2, 2, 3, 3, 4, 6, 6, 8, 10, 13, 14, 15, 18, 22, 25.

- (3p.) Oszacuj rozkład wysokości szkód, wykorzystując funkcję gęstości empirycznej (histogram). Przyjmij następujące granice przedziałów klasowych: 0.5, 2.5, 8.5, 15.5, 29.5.
- (2p.) Za pomocą otrzymanego modelu empirycznego, oszacuj dla tych szkód ograniczoną wartość oczekiwana (*limited expected value*):  $E(X \wedge 20)$ .

a) Emperyczna funkcja gęstości (histogram) przyjmuje postać:

$$f_m(x) = \frac{n_j}{n(c_j - c_{j-1})}, \quad c_{j-1} \leq x < c_j, \quad j = 1, \dots, k, \quad \text{gdzie}$$

$n_j$  - liczba obserwacji w danym przedziale

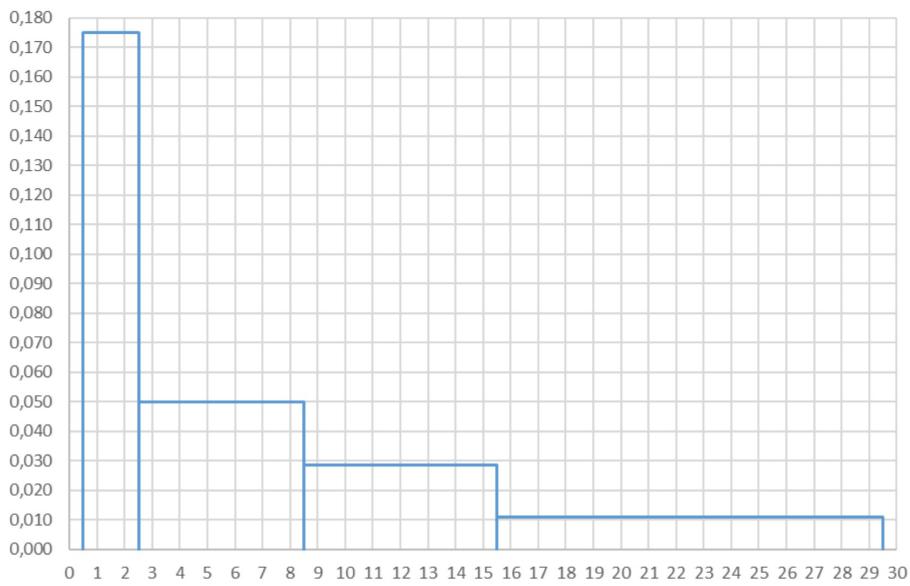
$n$  - łączna liczba obserwacji

$c_j - c_{j-1}$  - szerokość przedziału

$[c_{j-1}, c_j)$	$n_j$	$c_j - c_{j-1}$	$f_j$
$[0.5, 2.5)$	7	2	$\frac{7}{20 \cdot 2} = 0.1750$
$[2.5, 8.5)$	6	6	$\frac{6}{20 \cdot 6} = 0.0500$
$[8.5, 15.5)$	4	7	$\frac{4}{20 \cdot 7} = 0.0286$
$[15.5, 29.5)$	3	14	$\frac{3}{20 \cdot 14} = 0.0107$

$$f_m(x) = \begin{cases} 0.1750, & 0.5 \leq x < 2.5 \\ 0.0500, & 2.5 \leq x < 8.5 \\ 0.0286, & 8.5 \leq x < 15.5 \\ 0.0107, & 15.5 \leq x < 29.5 \end{cases}$$

Gęstość empiryczna

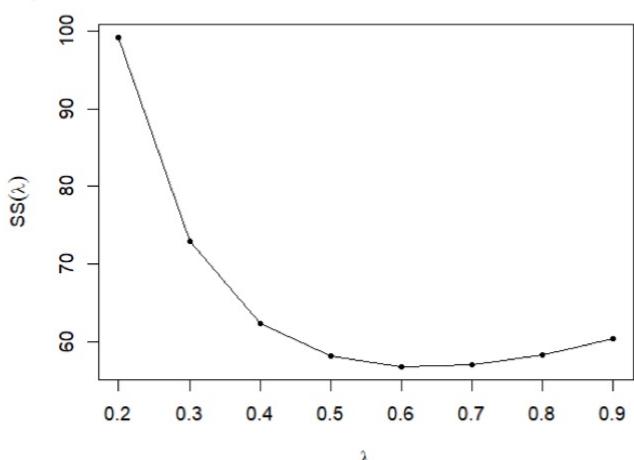


$$\begin{aligned}
 b) E[X|20] &= \int_{0.5}^{2.5} 0.175x \, dx + \int_{2.5}^{8.5} 0.05x \, dx + \int_{8.5}^{15.5} 0.0286x \, dx + \\
 &+ \int_{15.5}^{20} 0.0107x \, dx + \int_{20}^{29.5} 0.0107 \cdot 20 \, dx = \\
 &= 0.525 + 1.65 + 2.4024 + 0.2547 + 2.033 = \\
 &= 7.4651
 \end{aligned}$$

## Zadanie 5.

- a) (2p.) Czym jest wygładzanie wykładnicze (*exponential smoothing*) i jakie jest jego główne zastosowanie w analizie szeregów czasowych?
- b) (3p.) Zadaniem aktuariusza jest wyznaczenie prognozy pewnego szeregu czasowego  $y_t$ ,  $t = 1, \dots, 30$  na okres  $t = 31$ . W tym celu postanowił zastosować proste wygładzanie wykładnicze Browna. Aby ustalić optymalny parametr wygładzania  $\lambda$  wyznaczył sumę kwadratów błędów prognoz na jeden okres naprzód  $SS(\lambda)$  w zależności od  $\lambda$ . Uzyskane wyniki przedstawia rysunek 5.1

Rys. 5.1



Wybierz optymalną wartość parametru  $\lambda$  i wykorzystując informacje podane w tabeli 5.1 wyznacz prognozę na okres  $t = 31$ . Sprawdź dodatkowo jakość modelu wyznaczając błąd MAPE (*mean absolute percentage error*) dla prognoz na okresy  $t = 27, 28, 29, 30$ .

Tab. 5.1

Okres $t$	...	27	28	29	30
Wartość rzeczywista	...	12	16	15	17
Prognoza	...	13.354	12.542	14.617	

### Odp. a)

Wygładzanie wykładnicze to jedna z metod prognozowania i analizy szeregów czasowych, która nadaje większą wagę nowszym obserwacjom, jednocześnie stopniowo zmniejszając wpływ starszych danych. Jest to technika używana do wygładzania fluktuacji w danych, co ułatwia identyfikację trendów i wzorców.

Główne zastosowania:

- Prognozowanie krótkoterminowe.
- Wygładzanie danych – redukcja szumów w danych poprzez eliminowanie nagłych skoków i anomalii.
- Modelowanie trendów i sezonowości – bardziej zaawansowane wersje, jak podwójne i potrójne wygładzanie wykładnicze (Holt-Winters), pozwalają uwzględniać trend i sezonowość w danych.

Rodzaje wygładzania wykładniczego:

- Proste wygładzanie wykładnicze – stosowane do danych bez wyraźnego trendu ani sezonowości.
- Wygładzanie podwójne (Holt's method) – uwzględnia zarówno poziom, jak i trend.
- Wygładzanie potrójne (Holt-Winters method) – dodatkowo uwzględnia sezonowość.

Wygładzanie wykładnicze jest szczególnie cenione za swoją prostotę i skuteczność w prognozowaniu, zwłaszcza gdy dane wykazują krótkoterminowe fluktuacje.

b) Optymalny parametr myślnictwa to ten, który minimalizuje sumę kwadratów błędów prognoz. Z myślności wynika, iż  $\lambda = 0.6$ .

Prawe myślnictwo myślnicwe Browna:

$$y_{t+1}^P = \lambda y_t + (1-\lambda) y_t^P$$

Stąd:

$$y_{30}^P = 0.6 \cdot 15 + 0.4 \cdot 14.617 = 14.847$$

$$y_{31}^P = 0.6 \cdot 17 + 0.4 \cdot 14.847 = 16.139$$

Stąd MAPE:

$$\text{MAPE} = \frac{1}{n} \sum_{t=1}^n \left| \frac{y_t - y_t^P}{y_t} \right|$$

$$\text{MAPE} = \frac{1}{4} \left( \left| \frac{12 - 13.354}{12} \right| + \left| \frac{16 - 12.542}{16} \right| + \left| \frac{15 - 14.617}{15} \right| + \left| \frac{17 - 14.847}{17} \right| \right) = \\ = 0.1203$$

## Zadanie 6.

- a) (2p.) Czym jest interakcja w kontekście czynników ryzyka ubezpieczeniowego i dlaczego różni się od korelacji?
- b) (1p.) Liczbę szkód modelowano za pomocą regresji Poissona z linkiem kanonicznym. Uwzględniono dwie zmienne wyjaśniające:
- wiek kierowcy (zmienna `driver.age` w latach),
  - płeć kierowcy (zmienna `driver.gender` z dwiema kategoriami: Female i Male).

Oszacowano dwa modele:

- Model A, w którym **uwzględniono interakcję** między wiekiem i płcią kierowcy.
- Model B, w którym **nie uwzględniono interakcji** między wiekiem i płcią kierowcy.

Poniżej podano wyniki oszacowań odpowiednio modelu A i B:

### Model A:

```
glm(formula = clm.count ~ driver.gender + driver.age +
  driver.gender:driver.age, family = poisson())
```

Coefficients:

		Estimate	Std. Error	z value
(Intercept)	$\hat{\beta}_0$	-1.994709	0.622219	-3.206
driver.genderMale	$\hat{\beta}_1$	0.347212	0.662843	0.524
driver.age	$\hat{\beta}_2$	?	0.013917	0.350
driver.genderMale:driver.age	$\hat{\beta}_3$	?	0.014798	-0.603
---				
(Dispersion parameter for poisson family taken to be 1)				

### Model B

```
glm(formula = clm.count ~ driver.gender + driver.age, family =
  poisson())
```

Coefficients:

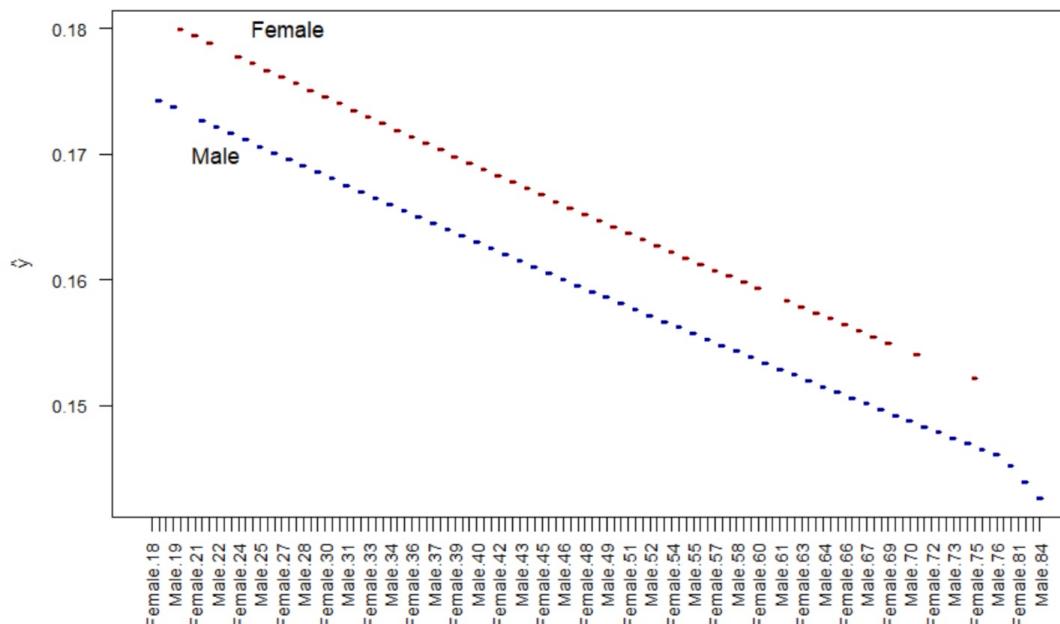
		Estimate	Std. Error	z value
(Intercept)		-1.654407	0.249672	-6.626
driver.genderMale		-0.037936	0.156919	-0.242
driver.age		-0.003044	0.004734	-0.643
---				

(Dispersion parameter for poisson family taken to be 1)

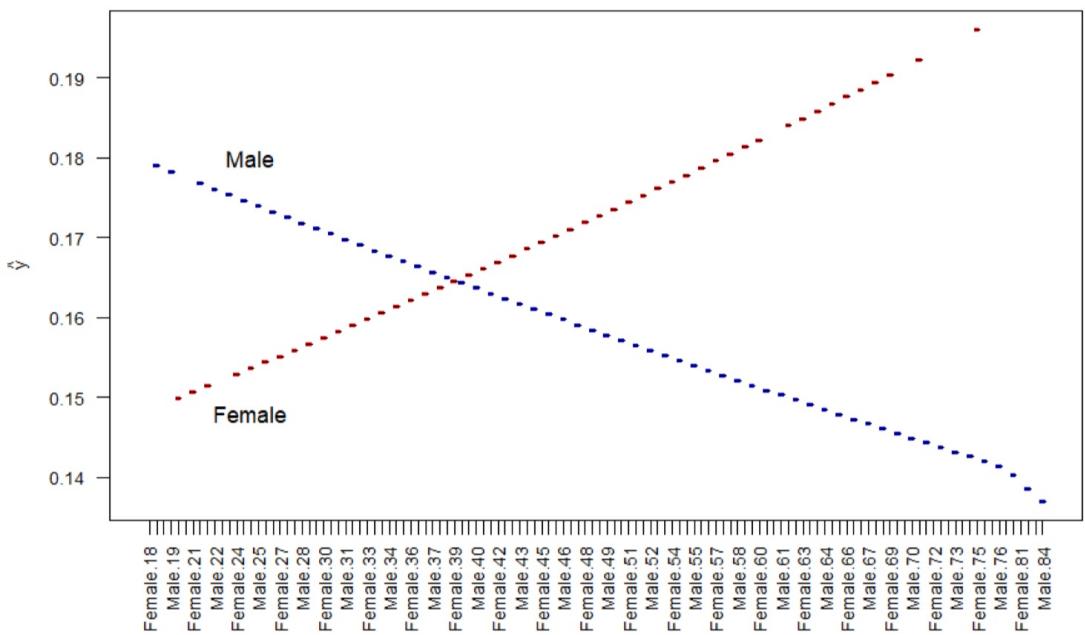
Wykorzystując obydwa modele wyznacz prognozę liczby szkód dla 30-sto letniego mężczyzny. Wiadomo, że dla modelu A:  $\hat{\beta}_2 + \hat{\beta}_3 = -0.004058$ .

- c) (2p.) Na rysunkach 6.1 i 6.2 zaprezentowano oszacowane wartości liczby szkód w zależności od wieku i płci. Wskaż, który rysunek odpowiada modelowi A, a który modelowi B, oraz uzasadnij swój wybór. Skomentuj wyniki przedstawione na tych rysunkach w kontekście, rozważanego w zadaniu, problemu interakcji.

Rys. 6.1



Rys. 6.2



### Odp. a)

Interakcja w kontekście czynników ryzyka ubezpieczeniowego odnosi się do sytuacji, w której wpływ jednego czynnika ryzyka na prawdopodobieństwo wystąpienia szkody zależy od obecności innego czynnika ryzyka. Oznacza to, że efekty dwóch (lub więcej) zmiennych nie są jedynie sumą ich indywidualnych wpływów, ale mogą się wzmacniać lub osłabiać w zależności od ich wzajemnego oddziaływaniania.

Różnica między interakcją a korelacją:

- Korelacja mierzy stopień współzmienności dwóch zmiennych, czyli na ile ich wartości są statystycznie powiązane (np. wzrost jednej zmiennej towarzyszy wzrostowi drugiej).
- Interakcja odnosi się do sposobu, w jaki jedna zmienna modyfikuje wpływ drugiej na określone zjawisko, np. ryzyko szkody. Może oznaczać, że efekt jednej zmiennej występuje tylko pod warunkiem obecności drugiej.

Korelacja sama w sobie nie oznacza interakcji – dwie zmienne mogą być silnie skorelowane, ale nie muszą wpływać na siebie nawzajem w sposób interakcyjny. Dlatego w analizie ryzyka ubezpieczeniowego ważne jest stosowanie modeli, które uwzględniają nie tylko współzależności, ale także potencjalne efekty interakcji między czynnikami.

### b) Model A

Przedykto r liniowy :

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{driver. gender} + \hat{\beta}_2 \cdot \text{driver. age} + \hat{\beta}_3 \cdot \text{driver. gender} \cdot \text{driver. age} =$$

$$= \begin{cases} \hat{\beta}_0 + \hat{\beta}_2 \cdot \text{driver. age} & , \text{ gdy driver. gender} = \text{Female} \\ \hat{\beta}_0 + \hat{\beta}_1 + (\hat{\beta}_2 + \hat{\beta}_3) \cdot \text{driver. age} & , \text{ gdy driver. gender} = \text{Male} \end{cases}$$

Prawdopodobieństwo w modelu Poissona :

$$\hat{y}_{30} = \exp[-1.9947 + 0.3472 - 0.004058 \cdot 30] = 0.170463$$

Model B:

$$\hat{y}_{30} = \exp \{-1.654407 - 0.037936 - 0.003044 \cdot 30\} = 0.170463$$

c) Rysunek 6.1 - Model B:

W modelu B różnicą między kobietami a mężczyznami to tylko czynnik  $-0.037936$ , oznacza to, że linie prognoz dla kobiet i mężczyzn ma stali logarytmicznej się różnoległo, czyli mają to samo nachylenie, równe współczynnikowi  $\hat{\beta}_2$  przy zmiennej driver. age i taka sytuacja pokazuje rysunek 6.1.

Automatycznie rysunek 6.2 odpowiada modelowi A.

Oba modele przedstawiają różne hipotezy na temat szkodowości.

Model B (Rys. 6.1) sugeruje prostą zależność: kobiety mają generalnie wyższą szkodowość niż mężczyźni, a dla obu płci ryzyko maleje w tym samym tempie wraz z wiekiem.

Model A (Rys. 6.2) opisuje bardziej złożoną rzeczywistość. Zgodnie z nim, młodzi mężczyźni są bardziej szkodowi niż młode kobiety. Jednak ich szkodowość gwałtownie spada z wiekiem (linia ma strome ujemne nachylenie). W przypadku kobiet szkodowość początkowo jest niższa, ale rośnie wraz z wiekiem (linia ma dodatnie nachylenie).

Problem interakcji polega na tym, że wpływ jednej zmiennej (np. wieku) na wynik jest różny na różnych poziomach innej zmiennej (np. płci). Wykres 6.2 jest klasycznym przykładem silnej interakcji – linie przecinają się, co oznacza, że w wieku ok. 36 lat szkodowość obu płci się zrównuje, a następnie relacja ryzyka się odwraca. Nieuwzględnienie interakcji (jak w modelu B) prowadzi do uproszczonego i potencjalnie błędnego opisu zjawiska.

## Zadanie 7.

- a) (3p.) Jakie są kluczowe różnice między krzywą Lorentza (*Lorenz curve*,  $LC[\hat{\mu}(X); \alpha]$ ) a krzywą koncentracji (*concentration curve*,  $CC[\mu(X), \hat{\mu}(X); \alpha]$ ) w kontekście oceny predyktora  $\hat{\mu}(X)$ .
- b) (2p.) W jaki sposób porównanie krzywej Lorentza i krzywej koncentracji pozwala ocenić jakość predyktora  $\hat{\mu}(X)$ ?

### Kluczowe różnice

- Mierzona wartość:
  - Krzywa Koncentracji ( $CC[\mu(X), \hat{\mu}(X); \alpha]$ ) mierzy skumulowany udział rzeczywistej składki (lub straty,  $\mu(X)$ ) dla  $\alpha$  procent polis o najniższych wartościach predyktora  $\hat{\mu}(X)$ . Innymi słowy, pokazuje, jaka część całkowitej szkody przypada na portfel uznany przez model za najmniej ryzykowny.
  - Krzywa Lorentza ( $LC[\hat{\mu}(X); \alpha]$ ) mierzy skumulowany udział prognozowanej składki  $\hat{\mu}(X)$  dla  $\alpha$  procent polis o najniższych wartościach tego samego predyktora  $\hat{\mu}(X)$ . Pokazuje, jaką część całkowitej zebranej składki stanowią polisy o najniższych prognozach.
- Cel i interpretacja:
  - Krzywa Koncentracji ocenia, jak dobrze uporządkowanie ryzyka przez predyktor odpowiada rzeczywistemu rozkładowi strat. Pokazuje, co powinno zostać zebrane z danego segmentu portfela.
  - Krzywa Lorenza opisuje wewnętrzną zmienność samego predyktora – jak bardzo różnicuje on składki. Pokazuje, co model proponuje zebrać z danego segmentu.

### Jak działa porównanie?

- Idealny predyktor: dla doskonałego predyktora, gdzie prognozowana składka jest równa składce rzeczywistej ( $\hat{\mu}(X) = \mu(X)$ ), obie krzywe pokrywają się.
- Ocena rozbieżności: różnica między krzywymi wskazuje na niedopasowanie modelu do ryzyka. Jeśli dla najmniej ryzykownych polis krzywa koncentracji leży powyżej krzywej Lorenza, oznacza to, że segment ten generuje większy udział w szkodach niż w zebranych składkach, co świadczy o ryzyku selekcji negatywnej (ang. adverse selection).

- Miara ilościowa: jakość predyktora można skwantyfikować za pomocą wskaźnika ABC (Area Between Curves), czyli pola powierzchni między krzywą koncentracji a krzywą Lorenza. Mniejsza wartość ABC oznacza, że struktura cenowa modelu jest bliższa rzeczywistej strukturze ryzyka, co świadczy o wyższej jakości predyktora.

### Zadanie 8.

- a) (2p.) Jaką rolę w wykrywaniu oszustw ubezpieczeniowych odgrywają techniki statystyczne, takie jak test chi-kwadrat czy analiza regresji?
- b) (3p.) Twoim zadaniem jest zbadanie możliwości wystąpienia oszustw w pewnym portfelu ubezpieczeń komunikacyjnych polegających na zgłoszaniu roszczeń dotyczących wypadków drogowych, w których zgłaszający nie brali udziału. Dysponujesz danymi dotyczącymi liczby zgłaszających roszczenia na jeden wypadek (w ramach jednego wypadku) oraz wzorcowym rozkładem prawdopodobieństwa liczby zgłaszających roszczenia dla wypadków, w których nie stwierdzono oszustw. Dane te zostały przedstawione poniżej (Tab. 8.1):

Tab. 8.1

Liczba zgłaszających roszczenia w ramach jednego wypadku	Zaobserwowana liczba wypadków	Rozkład wzorcowy
1	235	0.25
2	335	0.35
3	250	0.24
4	111	0.11
5	47	0.04
6+	22	0.01
Suma	1000	1.00

Wykorzystując test zgodności chi-kwadrat (na poziomie istotności 0.05), wskaż czy w analizowanym przypadku wystąpił problem z oszustwami:

- Podaj hipotezę zerową i alternatywną.
- Podaj wzór na statystykę testową i wskaż jaki ma rozkład.
- Podejmij odpowiednią decyzję odnośnie postawionej hipotezy zerowej. Co ta decyzja oznacza w kontekście badania możliwości wystąpienia oszustw w analizowanym portfelu?

#### Odp. a)

Techniki statystyczne, jak test chi-kwadrat i analiza regresji pomagają identyfikować podejrzane wzorce i anomalie w danych, co może świadczyć o nieuczciwych działaniach. Test chi-kwadrat pozwala na szybkie wykrycie nietypowych rozkładów lub zależności, podczas gdy analiza regresji umożliwia głębsze zrozumienie zmiennych wpływających na ryzyko oszustwa. Na przykład:

- Zastosowanie testu chi-kwadrat:
  - Analiza liczby zgłaszanych roszczeń w określonych kategoriach (np. rodzaj szkody, czas zgłoszenia) w celu identyfikacji nietypowych wzorców.
  - Porównanie rozkładu częstości zgłaszanych roszczeń w różnych segmentach klientów, aby wykryć nadreprezentację podejrzanych roszczeń.
  - Sprawdzenie, czy występuje istotna statystycznie różnica między częstotliwością podejrzanych roszczeń a ogólną liczbą zgłoszeń.
- Zastosowanie regresji:
  - Ocena wpływu czynników, takich jak wiek zgłaszającego, liczba wcześniejszych roszczeń, miejsce zamieszkania, typ polisy na prawdopodobieństwo wystąpienia oszustwa.
  - Budowa modeli predykcyjnych, które klasyfikują zgłoszenia jako podejrzane lub prawidłowe.
  - Analiza reszt, aby identyfikować odstające obserwacje, które mogą sugerować nietypowe i potencjalnie oszukańcze zachowania.

Techniki statystyczne są często stosowane w połączeniu z metodami eksploracji danych (data mining) i uczeniem maszynowym, co zwiększa skuteczność wykrywania oszustw.

b) Hipotezy:

$H_0$ : Rozkład liczb zgłoszających roszczenia na jeden wypadek jest zgodny z mowym rozkładem.

$H_1$ : Rozkład liczb zgłoszających roszczenia na jeden wypadek jest różny od rozkładu mowego.

Statystyka testowa  $\chi^2$ :

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i},$$

$O_i$  - liczba zgłoszeń do wypadków

$E_i$  - liczba oczekiwanych do wypadków

Obliczenia pomocnicze:

Liczba zgłoszających roszczenia w ramach jednego wypadku	Zaobserwowa liczba wypadków	Rozkład wzorcowy	$E_i$	$\frac{(E_i - O_i)^2}{E_i}$
1	235	0.25	250	0.900
2	335	0.35	350	0.643
3	250	0.24	240	0.417
4	111	0.11	110	0.009
5	47	0.04	40	1.225
6+	22	0.01	10	14.400
Suma	1000	1.00	1000	<b>17.594</b>

Statystyka testowa ma  $df = 6 - 1 = 5$  stopni swobody

Wartość krytyczna:

$$\chi^2_{kr} = \chi^2_{0.05; 5} = 11.070 \quad - \text{odczytane z tabelic}$$

Ponieważ  $\chi^2 > \chi^2_{kr}$  ( $17.594 > 11.070$ ) to odrzucajemy hipotezę  $H_0$

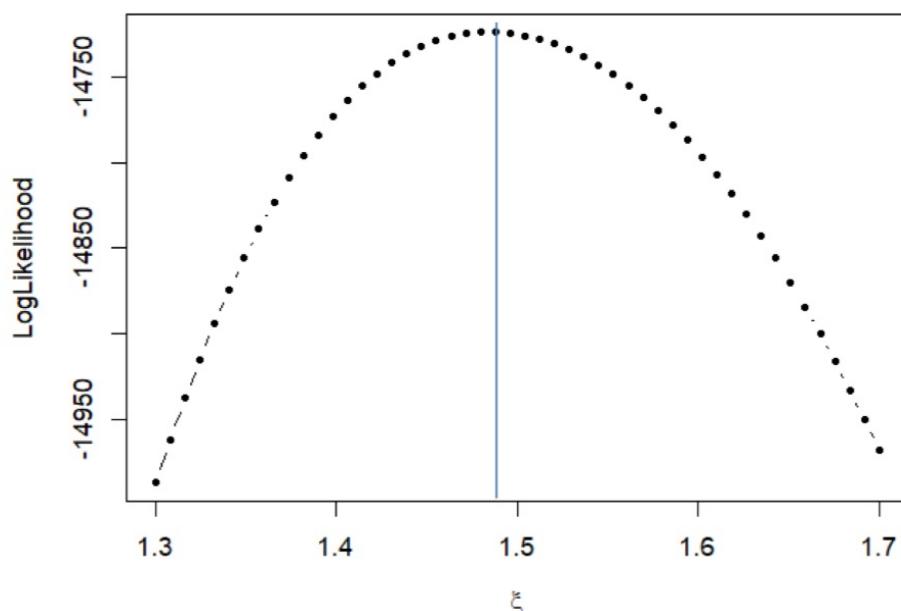
i przyjmujemy  $H_1$ . Oznacza to problem z oszustwem.

### Zadanie 9.

- a) (1p.) Jakie jest związek między wartością parametru  $\xi$  (*exponent parameter*) a stopniem heteroskedastyczności (zmienności) w modelach aktuarialnych opartych na rodzinie rozkładów Tweedie?
- b) (2p.) Omów zastosowanie rozkładu Tweedie w uogólnionych modelach liniowych (GLM) w kontekście modelowania roszczeń. Zwróć uwagę na rolę zmiennych objaśniających, które mogą wpływać zarówno na liczbę roszczeń, jak i na ich wysokość.
- c) (2p.) Aktuariusz zdecydował modelować składkę czystą (*pure premium*, całkowite roszczenie dla  $i$ -tej polisy przez jej ekspozycję) za pomocą uogólnionego modelu liniowego z wykorzystaniem rozkładu Tweedie i linku logarytmicznego. Uwzględnił trzy zmienne objaśniające:
- wiek samochodu (zmienna CarAge w latach),
  - rodzaj silnika (zmienna Gas z dwiema kategoriami: Diesel i Regular),
  - gęstość zaludnienia (zmienna Density w liczba osób/ km<sup>2</sup>).

W celu oszacowania tego modelu skonstruował przedstawiony na rys. 9.1 profil logarytmu wiarygodności (wartość logarytmu funkcji wiarygodności modelu w zależności od parametru  $\xi$  ).

Rys. 9.1



Następnie ustalił odpowiednią wartość parametru  $\xi$  i oszacował model. Uzyskał następujące wyniki:

Call:

```
glm(formula = Pure.Prem ~ CarAge + Gas + Density, family = tweedie(var.power = ?, link.power = 0))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	4.881e+00	1.141e-01	42.759	< 2e-16	***
CarAge	-2.122e-02	1.115e-02	-1.903	0.057065	.
GasRegular	-4.135e-01	1.202e-01	-3.441	0.000581	***
Density	2.246e-05	1.253e-05	1.793	0.072957	.
---					

(Dispersion parameter for Tweedie family taken to be **160.1988**)

Wskaż dla jakiej wartości parametru  $\xi$  aktuariusz oszacował model. Podaj oszacowania wartości oczekiwanej i wariancji zmiennej modelującej składkę czystą dla 7-mio letniego samochodu z silnikiem diesla, zarejestrowanego w regionie o gęstości zaludnienia 146.

### Odp. a)

W modelach aktuarialnych opartych na rodzinie rozkładów Tweedie, parametr  $\xi$  kontroluje zależność wariancji od wartości oczekiwanej, co bezpośrednio wpływa na heteroskedastyczność modelu. Związek ten opisuje równanie wariancji:

$$Var(Y) = \varphi\mu^\xi$$

gdzie:

$\varphi$  – parametr skali,

$\mu$  – wartość oczekiwana,

$\xi$  – parametr kształtu (*exponent parameter*).

Im większe  $\xi$ , tym silniejsza heteroskedastyczność, co oznacza, że wariancja bardziej zależy od wartości oczekiwanej.

b)

Rozkład Tweedie w ramach uogólnionych modeli liniowych (GLM) jest stosowany do modelowania łącznej kwoty roszczeń, która jest wypadkową zarówno ich liczby, jak i wysokości. Rozkłady te, dla parametru potęgowego (wykładnika)  $\xi$  z przedziału (1, 2), odpowiadają złożonemu rozkładowi Poissona z sumami o rozkładzie Gamma.

W tym podejściu:

- Liczba roszczeń (składnik częstotliwości) jest modelowana przy użyciu rozkładu Poissona.
- Wysokość pojedynczego roszczenia (składnik szkodowości) jest modelowana przy użyciu rozkładu Gamma.

### Rola zmiennych objaśniających i ograniczenia modelu

Głównym ograniczeniem standardowego modelu Tweedie GLM jest założenie o stałym parametrze dyspersji  $\phi$ . To założenie narzuca istotne ograniczenie na wpływ zmiennych objaśniających na składniki ryzyka:

- Każdy czynnik ryzyka (zmienna objaśniająca), który wpływa na oczekiwany wysokość roszczenia (składnik Gamma), musi również wpływać na oczekiwany liczbę roszczeń (składnik Poissona) w tym samym kierunku. Oznacza to, że jeśli dany czynnik zwiększa średnią szkodę, musi również zwiększać średnią częstotliwość roszczeń, aby parametr dyspersji  $\phi$  pozostał stały.

Ograniczenie to często jest sprzeczne z rzeczywistymi obserwacjami w ubezpieczeniach. Na przykład:

- W rezerwach szkodowych oczekiwana liczba wypłat zazwyczaj maleje w kolejnych latach rozwoju szkody, podczas gdy średnia kwota wypłaty rośnie.
- W ubezpieczeniach komunikacyjnych czynniki geograficzne mogą działać w przeciwnych kierunkach – w dużych miastach obserwuje się wyższą częstotliwość szkód, ale niższą ich średnią wysokość.

Ze względu na te przeciwnostawne tendencje, stosowanie modelu Tweedie z stałym parametrem dyspersji może prowadzić do zniekształcenia analizy.

c) Wartość parametru  $\xi$  trouba odnosiła się wynosiła  $\xi = 1,49$ .

Model w zadanym uogólniającej logarytmiczną funkcje Tylreca co oznacza, iż:  $\ln(\mu) = \gamma$ , a wtedy  $\mu = \exp\{\gamma\}$ .

$$\gamma = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{CarAge} + \hat{\beta}_2 \cdot \text{GasRegular} + \hat{\beta}_3 \cdot \text{Density}$$

$$\text{CarAge} = 7$$

$$\text{GasRegular} = 0 \quad \text{bo many diesel}$$

$$\text{Density} = 146$$

$$\hat{\mu} = \exp\{4.881 - 0.02122 \cdot 7 + 0.00002246 \cdot 146\} = \exp\{4.7357\} = 113.9432$$

Wariancja obliczana ze wzoru:

$$\widehat{\text{Var}}(y_i) = \varphi \mu^\xi$$

$$\widehat{\text{Var}}(y_i) = 160.1988 \cdot 113.9432^{1.49} = 185833.92$$

### Zadanie 10.

Poniższe pytania odnoszą się do hierarchicznego grupowania aglomeracyjnego (*bottom-up*).

- (1p.) Co oznacza wykonanie „poziomego cięcia” na dendrogramie?
- (1p.) Dlaczego wysokość połączenia na dendrogramie jest ważna przy interpretacji wyników grupowania?
- (3p.) Za pomocą hierarchicznego grupowania aglomeracyjnego analizowano zbiór ośmiu obiektów opisanych dwiema zmiennymi objaśniającymi  $X$  i  $Y$ . Przyjęto odległość Euklidesa oraz wiązanie najdalszego sąsiada (*complete linkage*). Odpowiedni zbiór danych podano w tabeli 10.1.

Tab. 10.1

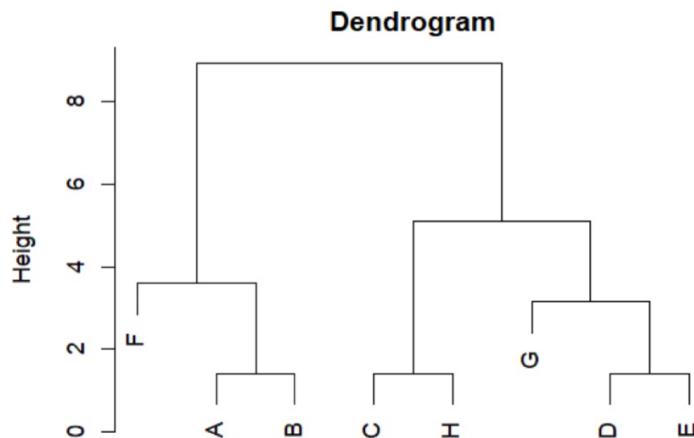
Obiekt	$x_i$	$y_i$
A	1	2
B	2	3
C	5	8
D	8	7
E	9	6
F	3	5
G	7	4
H	6	9

Na rysunkach 10.1 i 10.2 przedstawiono odpowiednio: otrzymany dendrogram i wyniki podziału na trzy skupienia:

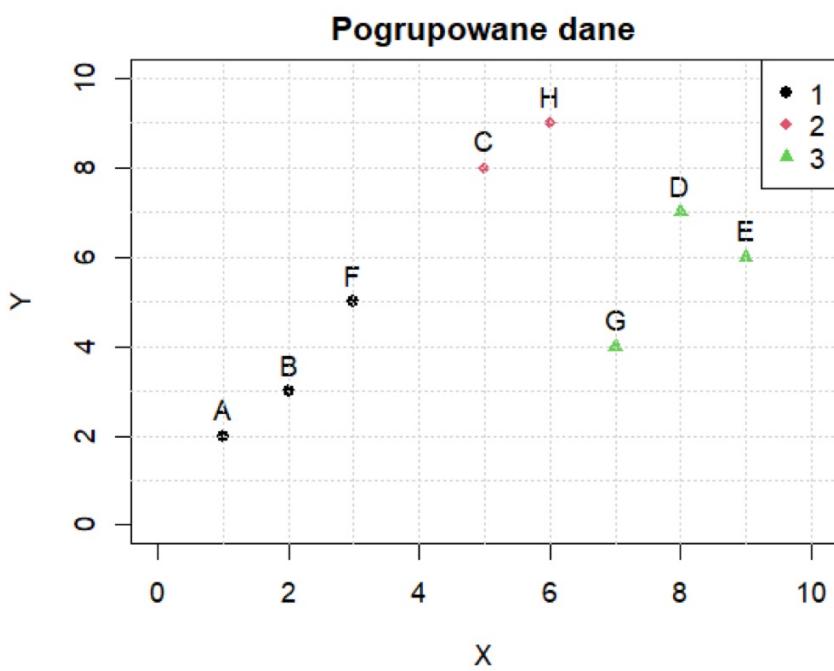
- 1: {A, B, F},
- 2: {C, H},
- 3: {D, E, G}.

Wyjaśnij w jaki sposób uzyskano taki podział oraz oblicz odległość między skupieniem 2 i 3. Zaznacz ją na dendrogramie przedstawionym na rysunku 10.1.

Rys. 10.1



Rys. 10.2



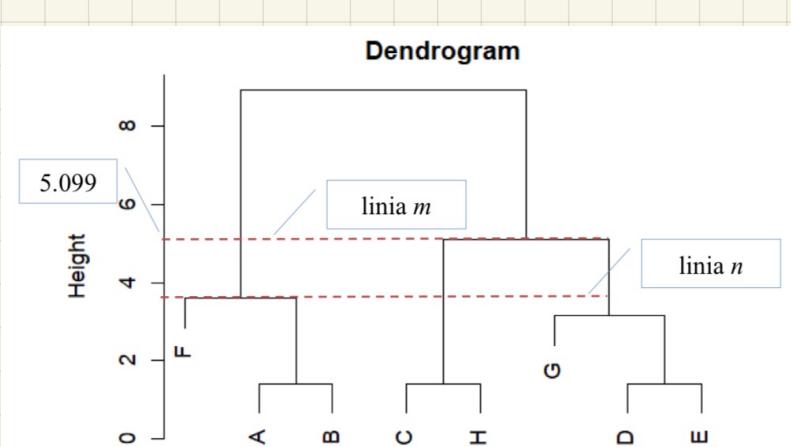
### Odp. a)

Dendrogram to drzewiasta struktura, która przedstawia sposób łączenia obiektów w grupy (skupienia, klastry) na różnych poziomach podobieństwa. Poziome cięcie na określonej wysokości oznacza usunięcie połączeń powyżej tej wartości, co prowadzi do podziału obiektów na skupienia. Ich liczba zależy od wysokości, na której wykonano cięcie – im niżzej, tym więcej mniejszych skupień, im wyżej, tym mniej większych.

### Odp. b)

Wysokość połączenia odzwierciedla odległość (niepodobieństwo) między grupowanymi obiektami lub skupieniami. Im wyżej następuje połączenie, tym większe różnice między skupieniami. Pomaga to wybrać odpowiednią liczbę skupień oraz zrozumieć strukturę danych.

c)



Kiedy przecięcią gatunki między liniami  $m$  i  $n$  powstaje grupka podział na trzy skupienia:  $\{A, B, F\}$ ,  $\{C, H\}$ ,  $\{D, E, G\}$ .

Zastosowano miernik najbliższego sąsiada. Oznacza to, że odległość między dwoma skupieniami jest równa minimalnej odległości Euclideanowej pomiędzy dowolnymi obiektami z obu skupień.

Najdalej położone od siebie punkty z grupy 2 i 3 to punkty G i H co widać na drugim wykresie:

$$d(G, H) = \sqrt{1^2 + 5^2} = 5.099$$