Zadanie 1.

Na podstawie danych z tabeli 1.1 oszacowano uogólniony model liniowy z linkiem kanonicznym, zakładając, że zmienna objaśniana Y podlega rozkładowi Poissona. Uzyskano następujące oszacowania parametrów: $\hat{\beta}_0 = \ln(3), \hat{\beta}_1 = \ln(1.5)$.

Tab. 1.1

x_i	y_i
0	2
0	2
0	3
0	5
1	3
1	4
1	5
1	6

a) (**2p.**) Oblicz dewiancję D tego modelu.

b) (**2p.**) Oblicz i zinterpretuj wartość $\frac{D}{df}$, gdzie df oznacza liczę stopni swobody reszt (*residua degrees of freedom*) dla tego modelu. Czy bardzo mała wartość $\frac{D}{df}$ jest zawsze dobrą wiadomością? Odpowiedź uzasadnij.

c) (**1p.**) Wyjaśnij, dlaczego dewiancja modelu nasyconego wynosi 0.

a) Dewiancja dla modelu regresji Poissona:

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

$$\hat{y}_i = \hat{\mu}_i = \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_i\} = \exp\{\ln(3) + \ln(1.5)\, x_i\}$$

Obliczenia pomocnicze:

x_i	y_i	$\hat{\mu}_i$	$y_i \ln\left(\frac{y_i}{\hat{y}_i}\right)$
0	2	3	-0.81093
0	2	3	-0.81093
0	3	3	0.00000
0	5	3	2.55413
1	3	4.5	-1.21640
1	4	4.5	-0.47113
1	5	4.5	0.52680
1	6	4.5	1.72609
		Razem	1.49764

$D = 2 \cdot 1.49764 = 2.99528$

b) $\frac{D}{df} = \frac{2.99528}{8-2} = 0.4992117$

df – liczba stopni swobody czyli n – liczba estymowanych parametrów

Jest to estymator dyspersji. W dobrze dopasowanym modelu Poissona oczekujemy $\frac{D}{df} \approx 1$. Bardzo mała wartość tego wskaźnika nie jest jednak dobrą wiadomością - może wskazywać na zbyt małe reszty, nadmierne dopasowanie (overfitting) lub błędną specyfikację modelu. W efekcie testy istotności mogą być zbyt optymistyczne, a wnioski – mylące.

Estymowana dyspersja mniejsza od 1 mówi o tym, że wariancja w dopasowanym modelu jest za niska, wartość większa od 1 mówi o zbyt dużej wariancji w dopasowanym modelu.

c)

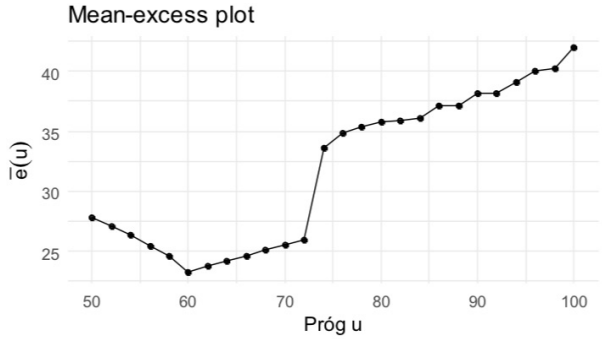
Model nasycony dokładnie odtwarza dane → nie ma żadnych reszt → brak „straty dopasowania" → *dewiancja* = 0.

Zadanie 2.

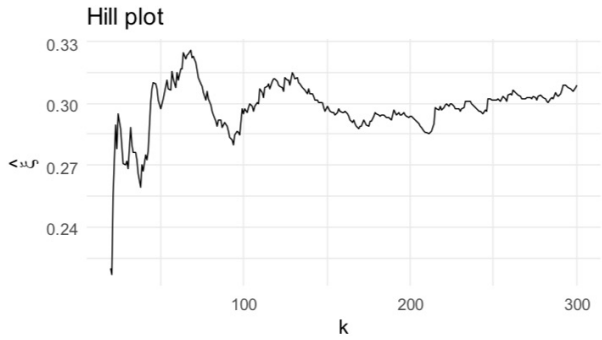
a) (**2*p.***) Wyjaśnij (odwołując się do odpowiedniego twierdzenia) dlaczego w modelu POT (*Peak Over Threshold*) nadwyżki $Y = X - u|X > u$ ponad wysoki próg u modeluje się uogólnionym rozkładem Pareto (*GPD*). Zinterpretuj parametr kształtu ξ (znak i konsekwencje dla „grubości" ogona) oraz podaj dwie praktyczne przesłanki wyboru progu u. Odpowiedź ogranicz do kilku precyzyjnych zdań.

b) (**3*p.***) Poniżej przedstawiono trzy wykresy diagnostyczne dla pewnego zbioru szkód (rys. 2.1-2.3)

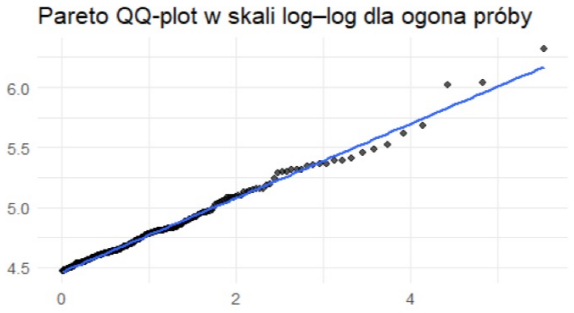
Rys. 2.1



Rys. 2.2



Rys. 2.3



- Wybierz konkretny próg u do modelu POT. Wybór uzasadnij w dwóch do trzech precyzyjnych zdaniach, odnosząc się do wykresów przedstawionych na rysunkach 2.1 i 2.2.
- Określ znak oraz przybliżony zakres wartości ξ na podstawie wykresów z rysunków 2.2 i 2.3. Jednym zdaniem wyjaśnij, co to oznacza dla tempa wzrostu wysokich kwantyli (np. $VaR_{0.995}$).

a)

Użycie uogólnionego rozkładu Pareto (GPD) do modelowania nadwyżki $Y = X - u|X > u$ ponad wysoki próg u w modelu POT (Peak Over Threshold) jest uzasadnione twierdzeniem Pickandsa-Balkemy-de Haana. Twierdzenie to wskazuje, że dla odpowiednio wysokich progów u, rozkład nadwyżek ponad ten próg może być dobrze przybliżony przez uogólniony rozkład Pareto.

Parametr kształtu ξ (nazywany również indeksem ogona lub indeksem wartości ekstremalnej) określa zachowanie ogona rozkładu:

- $\xi > 0$: odpowiada rozkładom o grubych ogonach (heavy-tailed). Wskazuje to, że ogon zanika wolno, jak w przypadku funkcji potęgowej.
- $\xi = 0$: granica wykładnicza, odpowiada rozkładom o lekkich ogonach.
- $\xi < 0$: Odpowiada rozkładom o ograniczonym z góry nośniku (krótkim ogonie).

Dwie praktyczne przesłanki przy wyborze progu u wynikają z kompromisu między obciążeniem a wariancją:

1. Na wykresie Mean-excess plot od danego rozpoczyna się stabilny, w przybliżeniu liniowy trend wzrostowy.
2. Na wykresie Hilla estymator parametru kształtu stabilizuje się w pewnym płaskim regionie, co wskazuje na znalezienie właściwego początku ogona rozkładu.

Odp. b)

- Wybór progu:
  - Na podstawie rys. 2.1 próg u można wskazać na poziomie 74-76. Wykres przechodzi w odcinek prawie liniowy o dodatnim nachyleniu — to sygnał, że nadwyżki powyżej u są dobrze przybliżane GPD.
  - Na podstawie rys. 2.2 próg u można wskazać na „półce" oscyluje wokół wartości około 0.29 ($k \approx 150$).
- Znak oraz przybliżony zakres wartości ξ:
  - $\xi > 0$ (ciężki ogon), $\xi \approx 0.28 - 0.31$
  - *Hill plot* (rys. 2.2) na „półce" oscyluje wokół wartości około 0.29, co sugeruje ξ rzędu 0.28–0.31. Z kolei, *Pareto QQ-plot (log–log)* (rys. 2.3) jest zbliżony do linii prostej o dodatnim nachyleniu w ogonie; dodatnie nachylenie i dobra liniowość na skali log–log są charakterystyczne dla ogonów Pareto-podobnych ($\xi > 0$).
- Konsekwencja dla wysokich kwantyli: Dla $\xi > 0$ wysokie kwantyle (np. VaR_α) rosną szybko w miarę zbliżania się α do 1, im większa ξ, tym szybszy wzrost.

Zadanie 3.

W pewnym zakładzie ubezpieczeń przeprowadzono kampanię promującą nowy produkt. Jej skuteczność („tak/nie") analizowano za pomocą losowo dobranej próby $n = 5000$ ubezpieczonych. Niech π_i oznacza prawdopodobieństwo, że i-ty ubezpieczony kupi nowy produkt. Te indywidualne prawdopodobieństwa modelowano za pomocą regresji logistycznej, wykorzystując dwie cechy objaśniające: wiek (*wiek* - zmienna ilościowa w latach) oraz region (*region* - kategorie: A, B, C).

a) (**2*p.***) Oszacowano następujący model **M1**:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.59980    0.07922 -32.816  < 2e-16 ***
wiek         0.19886    0.04997   3.979 6.91e-05 ***
regionB      0.05896    0.12383   0.476    0.634
regionC     -0.14887    0.15132  -0.984    0.325
---
    Null deviance: 2598.0  on 4999  degrees of freedom
Residual deviance: 2580.7  on 4996  degrees of freedom
AIC: 2588.7
```

- Podaj wiersz x_i macierzy X (modelu **M1**) dla 35-letniego ubezpieczonego z regionu B.
- Zinterpretuj oszacowane parametry.

b) (**3*p.***) Z kolei, oszacowano następujący model **M2**, w którym uwzględniono interakcję pomiędzy wiekiem a regionem:

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.58098    0.08060 -32.022  < 2e-16 ***
wiek            0.14843    0.07156   2.074  0.03807 *
regionB         0.08324    0.12759   0.652  0.51414
regionC        -0.41725    0.18391  -2.269  0.02328 *
wiek:regionB   -0.09634    0.11311  -0.852  0.39435
wiek:regionC    0.43374    0.13684   3.170  0.00153 **
---
    Null deviance: 2598.0  on 4999  degrees of freedom
Residual deviance: 2566.3  on 4994  degrees of freedom
AIC: 2578.3
```

- Sformułuj odpowiednie hipotezy H_0 i H_1 (werbalnie i matematycznie), aby sprawdzić, czy interakcja pomiędzy wiekiem a regionem ma wpływ na prawdopodobieństwo, że ubezpieczony nabędzie produkt.
- Zaproponuj test, którym – korzystając z podanych wyników z R – można wyznaczyć statystykę testową. Jaki (asymptotyczny) rozkład ma ta statystyka testowa?
- Oblicz statystykę testową i podaj decyzję testową na poziomie istotności $\alpha = 0.05$.

Odp. a)

Kodowanie w M1. Z tabeli współczynników: `(Intercept, wiek, regionB, regionC` wnika, że region A jest kategorią referencyjną, a `regionB, regionC` to zmienne zero–jedynkowe. Zmienna wiek jest ilościowa (w latach).

Wiersz macierzy X dla 35-latka z regionu B:

$$\boldsymbol{x}_i = [1,\ 35,\ 1,\ 0]$$

Interpretacja parametrów (M1):

- `wiek: 0.19886.` Każdy dodatkowy rok zwiększa log-szanse o 0.1989, czyli iloraz szans $exp(0.19886) \approx 1.22$ (ok. +22% na rok), przy ustalonym regionie. Współczynnik jest istotny.
- `regionB:0.05896.` W regionie B iloraz szans jest równy $exp(0.05896) \approx 1.06$. Szanse, że ubezpieczony z regionu B kupi nowy produkt są o ok. 6% większe niż w ubezpieczonego z regionu referencyjnego A (efekt jednak jest nieistotny, p ≈ 0.63).
- `regionC: -0.14887.` W regionie C iloraz szans jest równy $exp(-0.14887) \approx 0.86$. Szanse, że ubezpieczony z regionu C kupi nowy produkt są o ok. 14% niższe niż w ubezpieczonego z regionu referencyjnego A (efekt jednak jest nieistotny, p ≈ 0.33).

Odp. b)

Niech γ_B i γ_c oznaczają parametry w modelu M2 dotyczące interakcji wieku z odpowiednio regionem B i regionem C.

Hipotezy:

- H_0: brak interakcji wieku z regionem - parametry interakcji są równe 0, tj. $\gamma_B = \gamma_c = 0$
- H_1: co najmniej jedna interakcja jest różna od zera, tj. $(\gamma_B,\ \gamma_c) \neq (0, 0)$.

Proponowany test: Test ilorazu wiarygodności (LRT) oparty na różnicy dewiancji modeli zagnieżdżonych

Statystyka testowa:

$$T = D_{M1} - D_{M2}$$

$T \sim \chi^2_{\upsilon}$, gdzie stopnie swobody $\upsilon = df_{M1} - df_{M2} = liczba\ dodanych\ parametrów$

Wartość statystyki testowej:

$$T = 2580.7 - 2566.3 = 14.4, \quad \upsilon = 2$$

Wartość krytyczna testu (odczytana z tablicy d, dla $\alpha = 0.05$ i $\nu = 2$) wynosi 5.991.

Ponieważ $T > 5.991$, hipotezę H_0 odrzucamy.

Wniosek: Interakcja wiek × region istotnie poprawia dopasowanie — efekt wieku zależy od regionu.

Zadanie 4.

a) (**2*p.***) Wskaż, w których etapach procesu modelowania ryzyka: (i) *kalibracja parametrów*; (ii) *wycena/ustalanie składek*; (iii) *kalkulacja kapitału* (np. VaR/TVaR, SCR); (iv) *stress testy/scenariusze*, metoda Monte Carlo jest szczególnie użyteczna? Dla jednego ze wskazanych etapów podaj krótkie uzasadnienie (2–3 zdania), wskazując konkretny powód (np. złożoność modelu, brak formuł zamkniętych, zależności/ogony, nieliniowe wypłaty, itp.).

b) (**3*p.***) Opisz schemat estymacji metodą Monte Carlo składki $\pi = E[(S-d)^+]$ dla $S = \sum_{i=1}^{N} X_i$ (gdzie N – zmienna losowa dla liczby szkód w okresie, X_i – zmienne losowe dla wysokości szkód, d - próg):

- co losujemy i w jakiej kolejności,
- jak definiujemy estymator $\hat{\pi}$ i jak szacujemy jego niepewność,
- w jednym zdaniu wskaż, jak w praktyce uwzględnia się zależności między szkodami w tym schemacie.

a)

(i) Kalibracja parametrów modelu

Kalibracja to proces dopasowywania parametrów modelu statystycznego do danych historycznych w taki sposób, aby model jak najlepiej odzwierciedlał rzeczywistość. W przypadku prostych modeli parametry można estymować analitycznie (np. metodą największej wiarygodności). Jednak w złożonych modelach funkcja wiarygodności może być niemożliwa do bezpośredniej maksymalizacji.

Metoda Monte Carlo jest szczególnie użyteczna, gdy nie ma zamkniętego wzoru na wiarygodność albo model obejmuje zależności i ciężkie ogony (np. kopule, mieszaniny,cenzurowanie). Umożliwia symulacyjną estymację parametrów (np. MLE wspierane symulacją, indirect inference czy Approximate Bayesian Computation) nawet wtedy, gdy gęstość jest trudna do zapisania. Dzięki temu można dopasować realistyczne modele bez upraszczania założeń.

(ii) Wycena i ustalanie składek

Wycena produktów ubezpieczeniowych (ustalanie składki) wymaga oszacowania wartości oczekiwanej przyszłych strat (tzw. składki czystej). Dla prostych produktów można to zrobić analitycznie. Jednak wiele nowoczesnych produktów ma skomplikowaną strukturę, która uniemożliwia proste obliczenia.

Metoda Monte Carlo pozwala na symulację ogromnej liczby możliwych scenariuszy przyszłych strat. Dla każdego scenariusza obliczana jest wartość wypłaty z ubezpieczenia. Średnia arytmetyczna z tych wypłat jest estymatorem wartości oczekiwanej straty. Pozwala to na wycenę nawet najbardziej skomplikowanych instrumentów. Metoda sprawdza się doskonale przy wycenie produktów, których wartość zależy od wielu skorelowanych czynników ryzyka (np. stopy procentowe, inflacja, kursy walut) oraz zawiera opcje, franszyzy, limity i inne nieliniowości. Przykładem może być wycena obligacji katastroficznych (CAT bonds), gdzie wypłata zależy od wystąpienia i skali zdarzenia naturalnego (np. huraganu), którego modelowanie jest niezwykle złożone.

(iii) Kalkulacja kapitału (VaR/TVaR, SCR)

Instytucje finansowe muszą utrzymywać kapitał na pokrycie nieoczekiwanych strat. Miary ryzyka takie jak Value-at-Risk (VaR) i Tail Value-at-Risk (TVaR) są standardem w ocenie wymogów kapitałowych (np. SCR w Solvency II). Obliczenie tych miar wymaga znajomości rozkładu prawdopodobieństwa zagregowanych strat całego portfela, co jest trywialne tylko w teorii.

Agregacja wielu zależnych od siebie ryzyk i modelowanie ogonów rozkładu. Analityczne wyznaczenie rozkładu sumy wielu zależnych zmiennych losowych o różnych rozkładach (często z grubymi ogonami) jest praktycznie niemożliwe. Metoda Monte Carlo pozwala "brutalną siłą" obliczeniową zbudować ten rozkład i precyzyjnie zmierzyć ryzyko w jego ekstremalnych obszarach (ogonach), co jest kluczowe dla wymogów kapitałowych.

(iv) Stress Testy i Analiza Scenariuszy

Stress testy polegają na ocenie, jak portfel lub cała instytucja zachowa się w warunkach ekstremalnych, ale prawdopodobnych kryzysów (np. krach na giełdzie, gwałtowny wzrost inflacji, wielka powódź). Celem jest zrozumienie odporności na szoki, które wykraczają poza standardowe założenia modelu.

Stresy są z natury wieloczynnikowe (np. skok częstości + wysokie szkody + szok rynkowy) i mogą obejmować nieliniowe efekty oraz zależności pomiędzy liniami/rynkami. MC pozwala generować spójne scenariusze i rozkłady wyników pod narzuconymi szokami, porównać wpływ na wyniki, rezerwy, kapitał oraz przeprowadzić odwrotny test warunków skrajnych i analizy wrażliwości bez przebudowy całego modelu.

b)

1. Co losujemy i w jakiej kolejności

   Dla ścieżek $m = 1, \dots, n$:

   1. Liczba szkód: wylosuj $N^{(m)}$ z dopasowanego rozkładu (np. Poissona, ujemnego dwumianowego).
   2. Wysokość szkód: wylosuj niezależnie $X_1^{(m)}, \dots, X_{N^{(m)}}^{(m)}$ z dopasowanego rozkładu (np. lognormalnego, gamma, Pareto).
   3. Agregacja i wypłata: policz $S^{(m)} = \sum_{i=1}^{N^{(m)}} X_i^{(m)}$ oraz $Y^{(m)} = (S^{(m)} - d)^+$.

2. Niepewność

   - Estymator składki (MC):

$$\hat{\pi} = \frac{1}{n} \sum_{m=1}^{n} Y^{(m)}$$

- Niepewność:

  $SE(\hat{\pi}) = \frac{\hat{s}}{\sqrt{n}}$, gdzie $\hat{s}^2 = \frac{1}{n-1}\sum_{m=1}^{n}(Y^{(m)} - \hat{\pi})^2$

  Przedział ufności dla poziomu $1-\alpha$:

  $\hat{\pi} \pm u_{1-\alpha/2}\hat{s}$ ($u_{1-\alpha/2}$ - kwantyl rozkładu normalnego standardowego).

3. Zależności między szkodami w obrębie portfela, między liniami lub między N i X_i uwzględnia się na etapie generowania przez wspólny czynnik (*latent factor*, *common shock*) lub kopułę.

Zadanie 5.

a) (**1*p.***) Wyjaśnij, czym różni się pojęcie stacjonarności w sensie szerokim (*weak stationarity*) od stacjonarności w sensie ścisłym (*strict stationarity*).

b) (**2*p.***) Podaj wzór na autokowariancję procesu AR(1) i wyjaśnij, jak zmienia się wraz z opóźnieniem k.

c) (**2*p.***) Zadaniem aktuariusza jest wyznaczenie prognozy pewnego szeregu czasowego y_t, $t = 1, \ldots 30$ na okres $t = 31$. W tym celu postanowił zastosować proste wygładzanie wykładnicze Browna z parametrem wygładzania $\lambda = 0.6$. Wykorzystując informacje podane w tabeli 5.1:

- Sprawdź jakość modelu wyznaczając błąd MAPE (*mean absolute percentage error*) dla prognoz na okresy $t = 27, 28, 29, 30$.
- Wyznacz prognozę na okres $t = 31$.

Tab. 5.1

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

a)

Def. Stacjonarność w sensie ścisłym. Szereg czasowy $(X_t)_{t\in\mathbb{Z}}$ jest ściśle stacjonarny, jeśli

$$(X_{t_1}, \ldots, X_{t_n}) \overset{\mathrm{d}}{=} (X_{t_1+k}, \ldots, X_{t_n+k})$$

dla wszystkich $t_1, \ldots, t_n, k \in \mathbb{Z}$ oraz dla wszystkich $n \in \mathbb{N}$.

Def. Stacjonarność w sensie szerokim/słabym. Szereg czasowy $(X_t)_{t\in\mathbb{Z}}$ jest stacjonarny kowariancyjnie (lub słabo lub w sensie szerokim), jeśli jego dwa pierwsze momenty istnieją i spełniają warunki

$$\mu(t) = \mu, \quad t \in \mathbb{Z},$$

$$\gamma(t, s) = \gamma(t + k, s + k), \quad t, s, k \in \mathbb{Z}.$$

Szereg czasowy jest ściśle stacjonarny, jeśli jego łączny rozkład prawdopodobieństwa nie zmienia się w czasie. Oznacza to, że dla dowolnego zbioru punktów w czasie, łączny rozkład wartości w tych punktach jest taki sam, jak dla dowolnego innego zbioru punktów przesuniętych w czasie o stałą wartość.

Szereg czasowy jest stacjonarny w sensie szerokim, jeśli jego pierwsze dwa momenty (wartość oczekiwana i kowariancja) są stałe w czasie i skończone.

Podsumowując, stacjonarność ścisła jest silniejszym i bardziej ogólnym pojęciem, które gwarantuje niezmienność wszystkich charakterystyk statystycznych procesu w czasie. Stacjonarność w sensie szerokim jest słabszym warunkiem, ograniczającym się do stabilności wartości oczekiwanej, wariancji i autokowariancji, co w wielu praktycznych zastosowaniach jest wystarczające.

b)

Proces AR(1) jest zdefiniowany równaniem:

$$X_t = \phi_1 X_{t-1} + \epsilon_t$$

gdzie ϵ_t to proces białego szumu o wariancji σ_ϵ^2, a $|\phi_1| < 1$ jest warunkiem stacjonarności.

Wzór na funkcję autokowariancji $\gamma(k)$ dla opóźnienia (ang. *lag*) k jest następujący:

$$\gamma(k) = \frac{\phi_1^{|k|}\sigma_\epsilon^2}{1-\phi_1^2}$$

Wartość bezwzględna autokowariancji maleje wykładniczo. Ponieważ dla stacjonarnego procesu AR(1) musi zachodzić warunek $|\phi_1| < 1$, wartość $|\phi_1^{|k|}|$ maleje do zera w miarę wzrostu k. Oznacza to, że korelacja między odległymi w czasie obserwacjami jest coraz słabsza.

c)

Odp. c)

Proste wygładzanie wykładnicze Browna:

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

Stąd

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

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

Błąd MAPE:

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

$$= 0.1203$$

Zadanie 6.

a) (**1*p.***) Na czym polega model quasi-Poissona w uogólnionych modelach liniowych w porównaniu ze zwykłym modelem Poissona?

b) (**2*p.***) Wyjaśnij, jak w przypadku modelu quasi-Poissona należy interpretować wartość parametru dyspersji ϕ, w szczególności gdy:

- $\phi > 1$
- $0 < \phi < 1$.

Podaj przykłady przyczyn odpowiadających tym dwóm przypadkom (po jednej dla każdego).

c) (**2*p.***) Na podstawie tego samego zbioru danych oszacowano model Poissona i model quasi-Poissona. Uzyskano następujące wyniki:

Model Poissona

```
Call:
glm(formula = y ~ x, family = poisson)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.1787     0.2773   4.250 2.14e-05 ***
x             0.8362     0.3320   2.518   0.0118 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 53.262  on 7  degrees of freedom
Residual deviance: 46.354  on 6  degrees of freedom
```

Model quasi-Poissona

```
Call:
glm(formula = y ~ x, family = quasipoisson)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.1787          ?       ?       ?
x             0.8362          ?       ?       ?

(Dispersion parameter for quasipoisson family taken to be
7.144573)
```

Wypowiedz się na temat istotności oszacowanych parametrów w obu modelach (przyjmij poziom istotności równy 0.05).

a) Model quasi-Poissona w uogólnionych modelach liniowych działa podobnie jak zwykły model Poissona, ale dopuszcza, że wariancja nie musi być równa średniej. W praktyce oznacza to wprowadzenie parametru dyspersji, który pozwala modelować nadmierną lub zbyt małą zmienność danych (over- lub under-dispersion) bez zmiany postaci funkcji linku ani estymatorów średniej.

b)

W modelu quasi-Poissona parametr dyspersji ϕ określa, jak bardzo rzeczywista wariancja danych odbiega od tej zakładanej przez model Poissona.

- Gdy $\phi > 1$ - występuje nadmierna zmienność (overdispersion): wariancja danych jest większa niż średnia.

  Przykład przyczyny: nieobserwowana heterogeniczność między polisami. Część cech ryzyka nie została uwzględniona w modelu. W portfelu ubezpieczeń komunikacyjnych różni klienci mogą mieć odmienne zwyczaje jazdy, wiek pojazdu czy środowisko użytkowania, ale model może nie uwzględniać tych zmiennych. W efekcie częstotliwość szkód między jednostkami o pozornie tych samych cechach będzie bardziej zróżnicowana niż zakłada model.

- Gdy $0 < \phi < 1$ - występuje niedostateczna zmienność (underdispersion): dane są mniej zróżnicowane niż przewiduje model Poissona.

  Przykład przyczyny: agregacja dużej liczby podobnych jednostek ryzyka (np.portfele złożone z wielu małych, niezależnych polis), co „wygładza" zmienność obserwacji.

Odp. c)

W modelu Poissona oba parametry są statystycznie istotne.

Parametry modelu quasi-Poissona są takie same jak w modelu Poissona, ale ich błędy oszacowań SE^{qPois} są równe $\sqrt{\phi} \cdot SE^{Pois}$ (SE^{Pois} – błąd oszacowania parametru w modelu Poissona).
Z zamieszczonych wyników dla modelu quasi Poissona wynika, że $\sqrt{\phi} = 2.6729$.
Zatem błędy oszacowań SE^{qPois} są równe:

- `Intercept`: $0.2773 \cdot 2.6729 \approx 0.7412$
- `x`: $0.3320 \cdot 2.6729 \approx 0.88740$

Stąd statystyki t ($df = 6$, bo 8 obserwacji i 2 parametry)

- `Intercept`: $t = \frac{1.1787}{0.7412} \approx 1.590$
- `x`: $t = \frac{0.8362}{0.8874} \approx 0.942$

Wartość krytyczna dla testu dwustronnego na poziomie istotności równym 0.05 wynosi 2.447 (odczytana z tabeli e). Zatem nie ma podstaw do odrzucenia hipotez, że w modelu quasi Poissona zarówno wyraz wolny (`Intercept`), jak i parametr stojący przy zmiennej `x` są równe zero (są nieistotne).

Zadanie 7.

a) (**2p.**) Jakie jest znaczenie porównania krzywych $LC[\hat{\mu}(X);\alpha]$ (krzywa Lorenza) i $CC[\mu(X),\hat{\mu}(X);\alpha]$ (krzywa koncentracji) w ocenie adekwatności (sprawiedliwości) systemu taryf w ubezpieczeniach, tzn. zgodności składek $\hat{\mu}(X)$ z rzeczywistym kosztem ryzyka $\mu(X)$ w różnych grupach ryzyka?

b) (**3p.**) Dany jest portfel złożony z 5 polis (tab. 7.1):

Tab. 7.1

Polisa	Koszt $\mu(X)$	Składka $\hat{\mu}(X)$
1	100	80
2	200	150
3	300	250
4	200	300
5	200	220

Oblicz punkty LC i CC dla $\alpha = 0.4$ (dwie najmniejsze składki). Co uzyskane wyniki mówią o adekwatności taryfikacji w dolnych percentylach portfela?

a)

1. Diagnoza Adekwatności i Sprawiedliwości Taryfy

Idealna sytuacja to taka, w której obie krzywe niemal się pokrywają ($LC \approx CC$). Oznacza to, że składki $\hat{\mu}(X)$ są niemal idealnie proporcjonalne do rzeczywistego ryzyka $\mu(X)$.

Grupy ubezpieczonych, które wnoszą łącznie $\alpha\%$ całkowitej składki, generują również około $\alpha\%$ całkowitych szkód. Taki system taryfowy można uznać za adekwatny i sprawiedliwy z technicznego punktu widzenia, ponieważ każda grupa ryzyka płaci składkę odpowiadającą jej szkodowości. Brak jest wówczas systemowego subsydiowania jednych grup przez drugie.

2. Identyfikacja Kierunku i Skali Subsydiowania

Rozbieżność między krzywymi jest bezpośrednią miarą subsydiowania krzyżowego w portfelu.

Gdy LC > CC: Krzywa Lorenza (składek) znajduje się powyżej krzywej koncentracji (szkód). Oznacza to, że dla danego odsetka α "najtańszych" klientów, ich skumulowany udział w całkowitej składce jest większy niż ich skumulowany udział w szkodach. W efekcie, segmenty o niskim ryzyku nadpłacają w stosunku do generowanych przez siebie kosztów, subsydiując grupy o wyższym ryzyku.

Gdy CC > LC: Krzywa koncentracji jest powyżej krzywej Lorenza. Oznacza to, że skumulowane szkody dla segmentów o najniższych składkach rosną szybciej niż ich wkład w pulę składek. Te segmenty są subsydiowane – płacą za mało w stosunku do generowanego ryzyka.

Szerokość luki (odległość w pionie) między krzywymi wizualizuje skalę tego subsydiowania. Duża rozbieżność wskazuje na poważne niedopasowanie taryfy.

3. Weryfikacja Zgodności Rankingu Ryzyka

   Analiza ta odpowiada na fundamentalne pytanie: "Czy porządek w portfelu według wysokości składki $\hat{\mu}(X)$ jest zgodny z porządkiem według rzeczywistej szkodowości $\mu(X)$?".

   Trwałe i systematyczne odchylenia krzywych od siebie ujawniają błędy w strukturze taryfy. Na przykład, jeśli krzywa CC początkowo biegnie znacznie powyżej LC, a następnie ją przecina i biegnie poniżej, oznacza to, że taryfa jest "zbyt płaska". Klienci o niskim ryzyku płacą za dużo (względnie), a klienci o wysokim ryzyku za mało, ponieważ różnicowanie składek jest niewystarczające.

4. Lokalizacja Nieadekwatności w Portfelu

   Analiza krzywych pozwala precyzyjnie zidentyfikować, w których segmentach portfela (określonych przez parametr α) występuje największa rozbieżność.

   Jeśli krzywe znacząco się rozchodzą dla niskich wartości α, problem dotyczy segmentu klientów płacących najniższe składki.

   Jeśli rozbieżność pojawia się w środkowej części wykresu, nieadekwatność dotyczy "przeciętnych" klientów.

   Jeśli luki pojawiają się przy wysokich wartościach α, problem leży w segmencie klientów o najwyższym ryzyku i najwyższych składkach.

   Dzięki temu wiadomo, gdzie dokładnie należy wprowadzić korekty stawek.

5. Ocena Ryzyka Modelowego i Błędów Specyfikacji

   Stabilna, jednokierunkowa różnica między LC a CC jest silnym sygnałem, że model taryfowy $\hat{\mu}(X)$ cierpi na błąd specyfikacji. Nie jest to tylko kwestia drobnej korekty stawek, ale fundamentalnego problemu z modelem.

   Taka sytuacja sugeruje, że model może:

   - Pomijać istotne zmienne objaśniające ryzyko.
   - Używać nieprawidłowych wag lub offsetów (np. dla ekspozycji).
   - Mieć błędnie zdefiniowaną postać funkcyjną (np. liniową zamiast nieliniowej).

   Różnica LC–CC staje się więc miarą ryzyka modelowego, która wskazuje na konieczność rewizji i udoskonalenia samego modelu predykcyjnego, a nie tylko na prostą kalibrację stawek.

Odp. b)

Obliczenia dla $\alpha = 0.4$ (dwie najmniejsze składki)

Uporządkowanie po $\hat{\mu}$ (składkach)

Polisa	μ	$\hat{\mu}$
1	100	80
2	200	150
5	200	220
3	300	250
4	200	300

Suma składek = 1000; suma kosztów = 1000

Wartości LC i CC

$$LC[\hat{\mu}(X); 0.4] = \frac{230}{1000} = 0.23$$

$$CC[\mu(X), \hat{\mu}(X); 0.4] = \frac{300}{1000} = 0.30$$

Wśród 40% polis o najniższych składkach CC > LC (0,30 > 0,23). To oznacza, że ta grupa generuje 30% łącznego kosztu, a wnosi tylko 23% składek. W tej grupie polis składki są zbyt niskie i są subsydiowane przez resztę portfela. Taryfikacja w dolnych percentylach jest nieadekwatna.

$0.4 \cdot 5 = 2$ stąd dwie pierwsze polisy

$$LC = \frac{\text{Skumulowana składka}}{\text{Całkowita składka}} = \frac{230}{1000}$$

$$CC = \frac{\text{Skumulowany koszt}}{\text{Całkowity koszt}} = \frac{300}{1000}$$

Zadanie 8.

a) (**2*p.***) Dlaczego zgodnie ze standardami aktuariusz powinien analizować jakość i kompletność danych wykorzystywanych w kalibracji i walidacji modelu?

b) (**2*p.***) W portfelu brakuje pełnych danych o szkodach z lat 2019–2020. Aktuariusz chce jednak użyć modelu do wyznaczenia rezerw. Opisz dwa możliwe sposoby radzenia sobie z brakami danych.

c) (**1*p.***) Wyjaśnij, dlaczego transparentne udokumentowanie ograniczeń danych jest istotne w procesie sprawozdawczości i zarządzania ryzykiem w zakładzie ubezpieczeń.

a) Bez rzetelnych danych kalibracja i walidacja nie mają sensu.

- Błędy, niespójności i outliery zniekształcają estymację, prowadzą do błędnych parametrów i miar dopasowania.
- Braki, zmiany definicji lub niereprezentatywne okresy powodują obciążenie i niestabilność wyników (np. zawyżony VaR, złe stawki).
- Analiza jakości danych pozwala oszacować niepewność, przeprowadzić analizy wrażliwości i transparentnie komunikować ograniczenia.

Podsumowując, kontrola jakości i kompletności danych to warunek wiarygodności wyników i decyzji dotyczących np. taryfy, rezerwy, kapitału.

b) Na przykład:

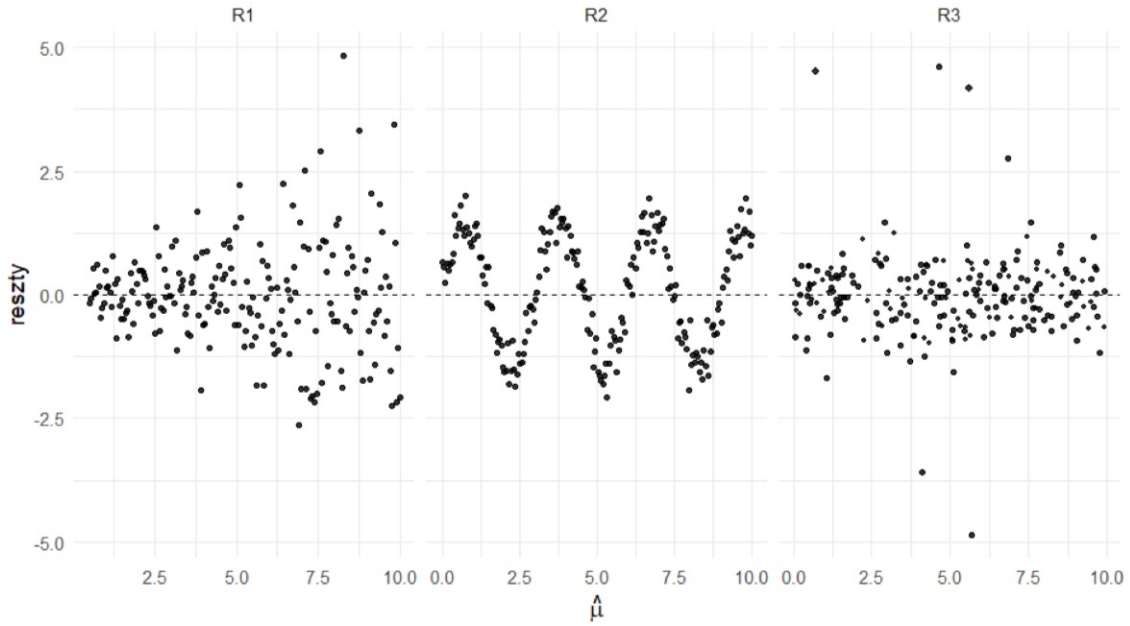
- Uzupełnienie braków i weryfikacja. Brakujące lata zastępujemy danymi porównywalnymi (z innych okresów/źródeł), korygując je o inflację, kalendarz szkód czy zmiany procesu likwidacji. Następnie sprawdzamy spójność i jakość(uzgodnienia, testy racjonalności, porównania historyczne), a wyniki prezentujemy wraz z analizą wrażliwości i jasnym opisem ograniczeń.
- Praca na tym, co jest, z marginesem ostrożności. Model kalibrujemy wyłącznie na dostępnych latach, a niepewność wynikającą z luk kompensujemy konserwatywnymi założeniami lub dodatkowymi marginesami. Efekt braków pokazujemy w krótkich scenariuszach/wrażliwościach i wyraźnie dokumentujemy wpływ na rezerwy.

c) Transparentne opisanie ograniczeń danych chroni przed „fałszywą precyzją" (pozwala właściwie odczytać wyniki, ich niepewność i zakres stosowalności modelu). Umożliwia świadome decyzje: wskazuje, gdzie wyniki są solidne, a gdzie wymagają ostrożności, marginesów lub scenariuszy. Podnosi przejrzystość i możliwość audytu, tzn. jasno opisana „ścieżka danych" (źródła, okresy, filtry, transformacje, imputacje), wraz z rejestrem podjętych decyzji i testów jakości, umożliwia odtworzenie obliczeń, zrozumienie zastosowanych korekt i ocenę ich wpływu na wyniki. Ułatwia też przeglądy wewnętrzne i kontrole zewnętrzne. Dzięki temu raporty są rzetelne, a zarządzanie ryzykiem bardziej odpowiedzialne i przejrzyste.

Zadanie 9.

a) (**1*p.***) Na poniższym rysunku (rys. 9.1) przedstawiono trzy wykresy (tj. R1, R2 i R3) reszt dewiancyjnych versus $\hat{\mu}$. Dopasuj każdy z nich do następujących diagnoz (i uzasadnij jednym zdaniem):
(A) - punkty wpływowe/outliers;
(B) - błędna funkcja wariancji/overdispersion;
(C) - brak istotnego predyktora lub złe przekształcenie.

Rys. 9.1



b) (**3*p.***) Zdefiniuj resztę Pearsona $r_i^{(P)}$ oraz standaryzowaną resztę Pearsona $\tilde{r}_i^{(P)}$. Wykorzystując model regresji Poissona (uogólniony model liniowy ze zmienną objaśnianą o rozkładzie Poissona) dla obserwacji $y_i = 10$ otrzymano $\hat{\mu}_i = 12$ oraz $h_i = 0.10$. Oblicz $r_i^{(P)}$ oraz $\tilde{r}_i^{(P)}$. Którą z tych reszt należy używać do porównań między obserwacjami i dlaczego?

c) (**1*p.***) Wyjaśnij, czemu w przypadku uogólnionego modelu liniowego ze zmienną objaśnianą o rozkładzie zero-jedynkowym, histogram reszt zwykłych bywa dwumodalny.

Odp. a)

- **(A) ↔R3:** Wykres z kilkoma punktami skrajnie oddalonymi (duże |reszty|), podczas gdy reszta chmury jest „prawidłowa" wokół 0.
- **(B) ↔R1:** Wykres o kształcie lejka: rozrzut reszt rośnie wraz z wartością dopasowaną ($\hat{\mu}$) (heteroscedastyczność).
- **(C) ↔R2** Wykres z systematycznym wzorcem (trend, łuk, U-kształ) wokół 0 zamiast losowej chmury, co świadczy o strukturze niewyjaśnionej przez model.

Odp. b)

Definicje:

- Reszta Pearsona:

$$r_i^{(P)} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)/\upsilon_i}}$$

- Standaryzowana reszta Pearsona:

$$\tilde{r}_i^{(P)} = \frac{r_i^{(P)}}{\sqrt{\hat{\phi}(1 - h_{ii})}},$$

gdzie h_{ii} to dźwignia (element diagonalny macierzy daszkowej).

Obliczenia dla modelu Poissona:

$$r_i^{(P)} = \frac{10 - 12}{\sqrt{12}} = -0.577$$

$$\tilde{r}_i^{(P)} = \frac{-0.577}{\sqrt{1 - 0.10}} = -0.609$$

Do porównań między obserwacjami lepsza jest standaryzowana reszta Pearsona, bo koryguje nie tylko heteroscedastyczność, ale też różnice dźwigni. Dzięki temu jej skala jest bardziej porównywalna w całym zbiorze.

Odp. c)

W GLM z rozkładem Bernoulliego (Binomialny z ($m = 1$)) zwykła reszta $e_i = y_i - \hat{\mu}_i$, gdzie $y_i \in \{0,1\}$ i $0 < \hat{\mu}_i < 1$.

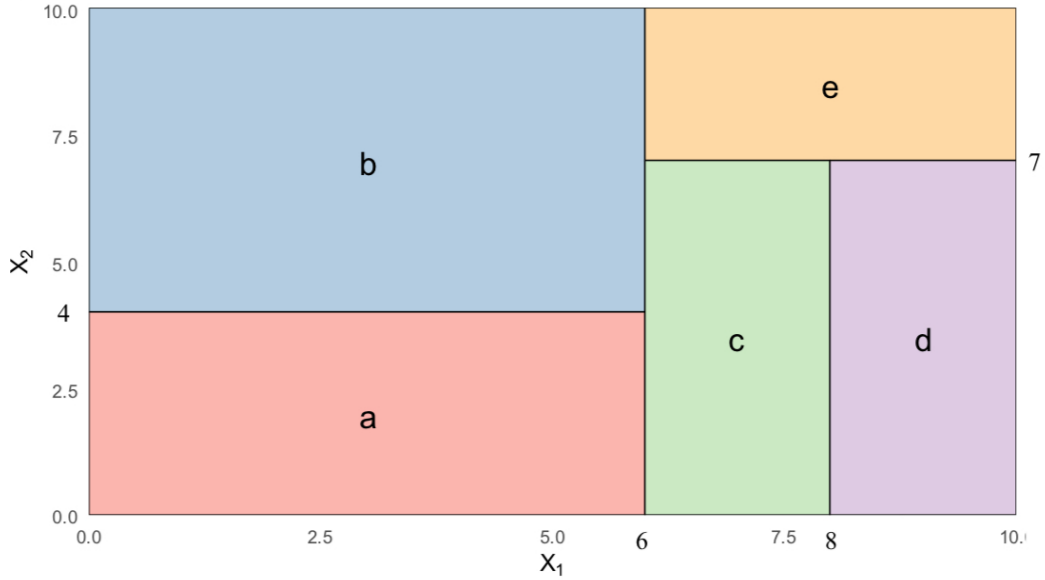
- Gdy $y_i = 1$, reszta wynosi: $1 - \hat{\mu}_i > 0$ (dodatnia).
- Gdy $y_i = 0$, reszta wynosi: $-\hat{\mu}_i$ (ujemna).

W całej próbie dostajemy więc mieszaninę dwóch „chmur" wartości: dodatnich i ujemnych, co naturalnie tworzy dwumodalny histogram. To powód, dla którego do diagnostyki w modelach binarnych preferuje się reszty Pearsona lub reszty dewiancyjne (standaryzowane), które mają bardziej symetryczne i porównywalne rozkłady.

Zadanie 10.

a) (**2*p.***) Poniżej (Rys.10.1) widzisz podział dwuwymiarowej przestrzeni cech utworzony przez klasyfikator typu CART**.** Narysuj odpowiadające mu drzewo decyzyjne.

Rys.10.1

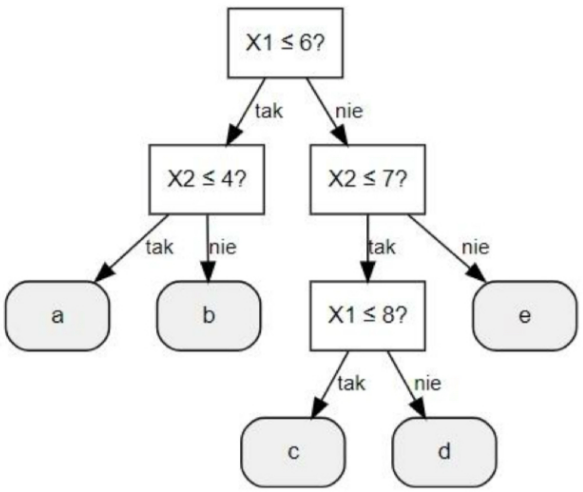


b) (**1*p.***) Wyjaśnij w jednym zdaniu, czym różni się proces uczenia kolejnych drzew w *baggingu* i *boostingu*.

c) (**1*p.***) Rozważ prostą regresję: rzeczywista wartość wynosi 100.
Model 1 (***bagging***): przewiduje 95
Model 2 (***boosting***): przewiduje 90, a następnie koryguje błąd +8 ($learning\ rate = 0.5$). Policz ostateczną predykcję dla boosting'u.

d) (**1*p.***) Która z tych metod (tj. *bagging*, *boosting*) lepiej radzi sobie z redukcją błędu systematycznego (*bias*)? Odpowiedź uzasadnij!

Odp. a)



.....

Odp. b)

Bagging uczy wiele drzew niezależnie i równolegle na bootstrapowych próbkach tych samych danych, a potem uśrednia ich przewidywania, natomiast *boosting* uczy drzewa sekwencyjnie, gdzie każde kolejne drzewo dopasowuje się do błędów poprzedniego zespołu (często z małym krokiem – *learning rate*) i koryguje dotychczasową prognozę.

Odp. c)

$\hat{y} = 90 + 0{,}5 \cdot 8 \;=\; 90 \;+\; 4 \;=\; 94.$

.....

Odp. d)

Odpowiedź: *Boosting*.

Boosting celuje w błąd systematyczny (bias) przez sekwencyjne korygowanie niedoszacowań/przeszacowań poprzednich modeli (uczenie na resztach lub wzdłuż gradientu straty), podczas gdy *bagging* przede wszystkim redukuje wariancję przez uśrednianie wielu niezależnie uczonych modeli.