Zadanie 1.

Liczbę szkód w pewnym portfelu ubezpieczeń AC modelowano z uwzględnieniem dwóch następujących zmiennych objaśniających:

- *veh_value* – wartość samochodu (zmienna ilościowa),
- *area* – zmienna jakościowa przyjmująca 6 kategorii: A, B, C, D, E i F określających jednostki administracji terytorialnej, w których zamieszkują właściciele samochodów.

Wykorzystując zbiór uczący oszacowano model regresji Poissona (z linkiem kanonicznym). Uzyskano następujące wyniki:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.954499   0.039988 -48.878  < 2e-16 ***
veh_value    0.054988   0.012447   4.418 9.97e-06 ***
areaB        0.009448   0.048450   0.195   0.8454
areaC       -0.011357   0.043899  -0.259   0.7959
areaD       -0.151883   0.059249  -2.563   0.0104 *
areaE       -0.087677   0.065028  -1.348   0.1776
areaF        0.082661   0.072558   1.139   0.2546
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Poisson family taken to be 1)

    Null deviance: 19396  on 47582  degrees of freedom
Residual deviance: 19365  on 47576  degrees of freedom
AIC: 26710
-----------
Analysis of Deviance Table
Terms added sequentially (first to last)

          Df Deviance Resid. Df Resid. Dev      Pr(>Chi)
NULL                     47582      19396.000
veh_value  1    18       47581      19378.000  1.644e-05 ***
area       5     a       47576      19365.167    b
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
-----------
```

a) ***(1p.)*** Podaj liczebność zbioru uczącego. Odpowiedź uzasadnij!

b) ***(1p.)*** Wyznacz wartości **a** i **b** w tabeli przedstawiającej wyniki analizy dewiancji.

c) ***(3p.)*** Na podstawie tego samego zbioru danych oszacowano nowy model regresji Poissona (z linkiem kanonicznym), w którym jako zmienne objaśniające wykorzystano dwie zmienne z wcześniejszego modelu oraz dodatkowo jedną zmienną ilościową. Uzyskano m.in. następujące wyniki:

```
    Null deviance: 19396  on 47582  degrees of freedom
Residual deviance: 19355  on 47575  degrees of freedom
AIC: 26701
```

Zbadaj, czy dodatkowa zmienna powinna zostać uwzględniona w modelu. W uzasadnieniu powołaj się m.in. na wyniki odpowiedniego testu na poziomie istotności 0.05.

a) 47583

Można to odczytać z wiersza „Null deviance" w podsumowaniu modelu.

Liczba stopni swobody dla deviancji zerowej jest obliczana jako $n-1$, gdzie n to liczba obserwacji w zbiorze danych.

$n - 1 = 47582$

$n = 47583$

b) Deviancja resztowa przed dodaniem „area": 19378.000

Deviancja resztowa po dodaniu „area": 19365.167

Zmiana deviancji:

$a = 19378.000 - 19365.167 = 12.833$

Wartość b należy odczytać z tablicy, na podstawie deviancji 12.833 i 5 stopni swobody:

$b = P(\chi^2_5 > 12.833) = 0.025$

c) Jak powinna. Wskazuje na to mniejsza wartość AIC dla modelu z dodatkową zmienną oraz wyniki testu ilorazu wiarygodności.

H_0: współczynnik dla nowej zmiennej jest równy zero

H_1: współczynnik dla nowej zmiennej jest różny od zera

Zmiana deviancji wynosi: $19365.167 - 19355 = 10.167$,

jest to wartość statystyki testowej:

$\chi^2 = 10.167$

Liczba stopni swobody: $47576 - 47575 = 1$

$10.167 > \chi^2_{0.05,1} = 3.841$

Odrzucamy H_0 i przyjmujemy H_1.

Zadanie 2.

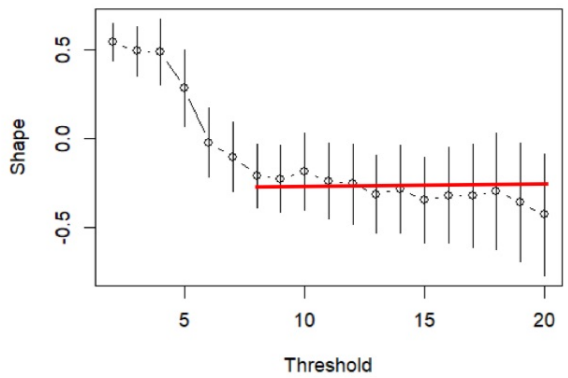
Poniższe pytania dotyczą zastosowania analizy przekroczeń progu (POT, *Peaks Over Threshold*) w badaniu zjawisk ekstremalnych.

a) (***2p.***) Dlaczego wybór odpowiedniego progu u jest kluczowy dla uzyskania trafnej aproksymacji przy użyciu uogólnionego rozkładu Pareto? (Jakie są konsekwencje wyboru zbyt niskiego lub zbyt wysokiego progu u?)

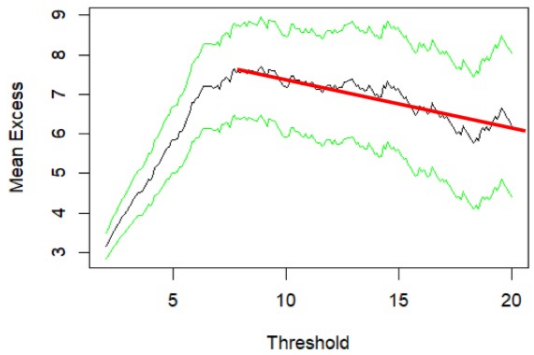
b) (***3p.***) Zadaniem aktuariusza jest oszacowanie kwantyla rzędu 0.995 dla wysokości roszczeń w określonym portfelu ubezpieczeń. Zdecydował się na zastosowanie metody POT, tworząc wykresy przedstawione na rysunku 2.1. Następnie oszacował uogólnione rozkłady Pareto dla progów $u = 4$ oraz $u = 8$. Wykresy kwantyl-kwantyl dla dopasowanych rozkładów zostały zaprezentowane na rysunku 2.2. Który próg powinien wybrać aktuariusz, aby uzyskać wiarygodne oszacowanie kwantyla? Odpowiedź uzasadnij, odwołując się do analizy **wszystkich** wykresów z rysunków 2.1 i 2.2.

Rys. 2.1

Pareto index plot

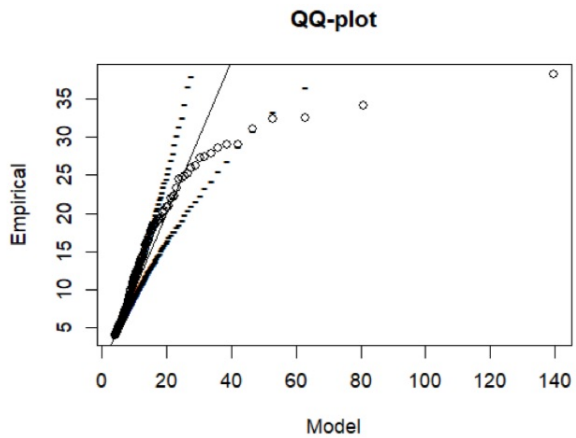


Mean residual life plot

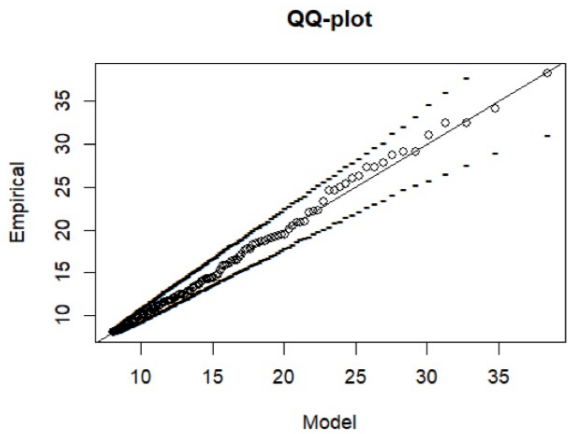


Rys. 2.2

a) $u = 4$



b) $u = 8$



a)

Wybór odpowiedniego progu `u` jest kluczowy, ponieważ wiąże się z fundamentalnym kompromisem między **obciążeniem** (błędem systematycznym) a **wariancją** (niestabilnością) oszacowań parametrów.

Konsekwencje niewłaściwego wyboru progu są następujące:

- **Wybór zbyt wysokiego progu `u`**: Powoduje, że do analizy trafia bardzo mało obserwacji (nadwyżek ponad próg). Prowadzi to do oszacowań o dużej wariancji, co oznacza, że są one **niestabilne i mało precyzyjne**.
- **Wybór zbyt niskiego progu `u`**: Skutkuje włączeniem do analizy obserwacji, które nie pochodzą z ekstremalnego ogona rozkładu. Dla tych danych założenie o uogólnionym rozkładzie Pareto (GPD) może nie być spełnione, co prowadzi do uzyskania **obciążonych, czyli systematycznie błędnych, oszacowań**.

Celem jest znalezienie najniższego możliwego progu, powyżej którego model GPD staje się trafną aproksymacją zachowania danych w ogonie rozkładu.

b) 1. Pareto index plot: ten wykres pokazuje jak zmienia się oszacowanie parametru kształtu w zależności od wybranego progu.
Od progu $u = 8$, wykres stabilizuje się, a estymacje parametru kształtu utrzymują się na stałym poziomie.

2. Mean residual life plot: ten wykres przedstawia średnią wartość przekroczenia ponad dany próg. Dla danych, które dobrze opisuje uogólniony rozkład Pareto, wykres ten powinien być w przybliżeniu liniowy powyżej właściwego progu.
Na wykresie widać, że dla progów powyżej $u = 8$ punkty układają się wzdłuż opadającej linii prostej.

3. Wykres kwantyl-kwantyl (QQ-plot) służy do oceny jakości dopasowania modelu teoretycznego do danych empirycznych.
Dla $u = 8$ dopasowanie modelu jest bardzo dobre. Punkty empiryczne leżą blisko linii prostej i mieszczą się wewnątrz przedziałów ufności.

Zadanie 3.

Posiadasz losową próbę pięciu obserwacji przedstawiających wysokości roszczeń ubezpieczeniowych. Trzy z tych wartości są znane i wynoszą odpowiednio 3, 6 i 14, natomiast dwie pozostałe są nieznane, ale wiadomo, że przekraczają kwotę 25.

a) ***(1p.)*** Czy posiadasz obserwacje ucięte (*truncated data*) czy cenzurowane (*censored data*)? Uzasadnij swoją odpowiedź.

b) ***(1p.)*** Opisz przynajmniej jeden przykład sytuacji, w której aktuariusz może spotkać się z tego rodzaju danymi.

c) ***(3p.)*** Wiadomo, że wysokości tych roszczeń podlegają rozkładowi o następującej funkcji gęstości:

$$f(x) = \frac{\alpha}{x^{(\alpha+1)}}, x > 1, 0 < \alpha < \infty.$$

Dysponując opisaną próbą losową, metodą największej wiarygodności oszacuj parametr α tego rozkładu.

a) Posiadamy dane cenzurowane

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

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

b) **Przykłady** Jednym z najczęstszych przykładów sytuacji, w której aktuariusz spotyka się z danymi cenzurowanymi z prawej strony, jest analiza szkód z polis ubezpieczeniowych, które posiadają limit odpowiedzialności (ang. policy limit).

W takiej sytuacji, jeśli rzeczywista wysokość szkody jest niższa od limitu, znamy jej dokładną wartość. Jednakże, jeśli szkoda przekracza limit, w danych ubezpieczyciela często zapisana jest jedynie kwota wypłaconego świadczenia, czyli wartość limitu. Oznacza to, że znamy tylko dolne ograniczenie rzeczywistej wysokości szkody (wiemy, że wyniosła co najmniej tyle, ile limit), ale jej dokładna wartość jest nieznana, czyli ocenzurowana.

Innym przykładem są badania śmiertelności. Jeśli badanie kończy się, a niektóre osoby wciąż żyją, ich dokładny wiek śmierci jest nieznany – wiemy tylko, że nastąpi on w przyszłości. Ich obserwacja jest więc cenzurowana w momencie zakończenia badania.

c) Dystrybuanta:

$$F(x) = \int_1^x \frac{\alpha}{t^{(\alpha+1)}}\, dt = -\frac{1}{t^{\alpha}}\Big|_1^x = 1 - \frac{1}{x^{\alpha}}$$

Funkcja przeżycia:

$$S(x) = 1 - F(x) = \frac{1}{x^{\alpha}}$$

Funkcja wiarygodności

$$L(\alpha) = \underbrace{f(3)\cdot f(6)\cdot f(14)}_{\text{obserwacje nie cenzurowane}} \cdot \underbrace{S(25)\cdot S(25)}_{\text{obserwacje cenzurowane}}$$

$$L(\alpha) = \alpha 3^{-\alpha-1} \cdot \alpha 6^{-\alpha-1} \cdot \alpha 14^{-\alpha-1} \cdot 25^{-2\alpha} =$$

$$= \alpha^3 \cdot \frac{1}{3} \cdot \frac{1}{6} \cdot \frac{1}{14} \left(3 \cdot 6 \cdot 14 \cdot 25^2\right)^{-\alpha}$$

$$\ln[L(\alpha)] = 3\ln(\alpha) + \ln\left(\frac{1}{3}\cdot\frac{1}{6}\cdot\frac{1}{14}\right) - \alpha \ln(157500)$$

$$\ln'[L(\alpha)] = 3\alpha^{-1} - \ln(157500) := 0$$

$$3\alpha^{-1} = \ln(157500)$$

$$\hat{\alpha} = \frac{3}{\ln(157500)} = 0.2507$$

Zadanie 4.

a) (**3*p.***) Wyjaśnij koncepcję oraz sposób konstrukcji wykresów PDP (*Partial Dependence Plot*).

b) (**2*p.***) Oszacowano liniowy model regresji przedstawiający zależność wysokości roszczeń (y) od wieku samochodu (*CarAge* - w latach) oraz gęstości zaludnienia w miejscowości użytkowania samochodu (*Density* – w ludność na 1km^2). Wykorzystano dane przedstawione w tabeli 4.1. Uzyskano następujący model:

$$\hat{y} = 49 - 6 \cdot CarAge - 0.01 \cdot Density.$$

Odpowiednie wartości dopasowane przedstawiono w ostatniej kolumnie tabeli 4.1. Na podstawie uzyskanych wyników skonstruowano wykres PDP dla zmiennej wiek samochodu (*CarAge*). Wyznacz wartość, która na tym wykresie odpowiada dwuletniemu samochodowi, tzn. wyznacz $\hat{y}_{PDP}$ dla $CarAge = 2$.

Tab. 4.1

Nr	y	*CarAge*	*Density*	$\hat{y}$
1	40	2	20	36.8
2	35	2	30	36.7
3	40	3	100	30
4	20	5	100	18
5	15	4	1000	15
6	10	5	1000	9

a)

Głównym problemem w interpretacji złożonych modeli jest trudność w zwizualizowaniu funkcji predykcyjnej, gdy zależy ona od więcej niż dwóch zmiennych (cech). Wykresy częściowej zależności (Partial Dependence Plots) rozwiązują ten problem, pokazując **marginalny wpływ** wybranego, małego podzbioru cech na wynik predykcji modelu.

Koncepcja polega na uśrednieniu wpływu wszystkich pozostałych cech, co pozwala na wyizolowanie i graficzne przedstawienie zależności między analizowaną cechą a prognozą. Innymi słowy, wykres pokazuje, jak zmieniłaby się średnia predykcja modelu, gdybyśmy zmieniali wartość jednej cechy, zachowując jednocześnie rozkład pozostałych cech z całego zbioru danych.

Sposób Konstrukcji

Konstrukcja wykresu PDP dla jednej wybranej cechy x_S odbywa się poprzez następujące kroki, które stanowią estymację funkcji częściowej zależności na podstawie zbioru treningowego:

1. **Wybierz cechę**, której wpływ chcesz zbadać (np. wiek kierowcy).
2. **Wybierz konkretną wartość** dla tej cechy (np. wiek = 30 lat).
3. Dla **każdej obserwacji** w zbiorze danych:
   - Zastąp oryginalną wartość wybranej cechy ustaloną wartością (np. dla każdego kierowcy w zbiorze ustaw wiek na 30 lat).
   - Pozostaw wartości wszystkich pozostałych cech bez zmian.
4. Dla każdej tak zmodyfikowanej obserwacji **wygeneruj predykcję** za pomocą modelu.
5. **Uśrednij wszystkie uzyskane predykcje**. Wynik to pojedynczy punkt na wykresie częściowej zależności dla wartości "wiek = 30 lat".
6. **Powtórz kroki 2-5** dla całego zakresu interesujących wartości wybranej cechy (np. dla wieku od 18 do 90 lat), aby uzyskać pełny wykres.

W ten sposób wykres pokazuje, jak zmienia się średnia predykcja modelu w zależności od wartości jednej cechy, po uśrednieniu efektów wszystkich innych zmiennych.

b) Należy obliczyć średnią prognozę modelu dla wszystkich obserwacji w zbiorze danych, podstawiając do wzoru CarAge = 2 i zachowując oryginalne wartości zmiennej Density z każdej obserwacji.

$\hat{y} = 49 - 6 \cdot CarAge - 0.01 \cdot Density$

$CarAge = 2$

$Density = 20, 30, 100, 100, 1000, 1000$

Nr	*CarAge*	*Density*	$\hat{y}(2, Density)$
1	2	20	36.8
2	2	30	36.7
3	2	100	36.0
4	2	100	36.0
5	2	1000	27.0
6	2	1000	27.0
		Σ	199.5

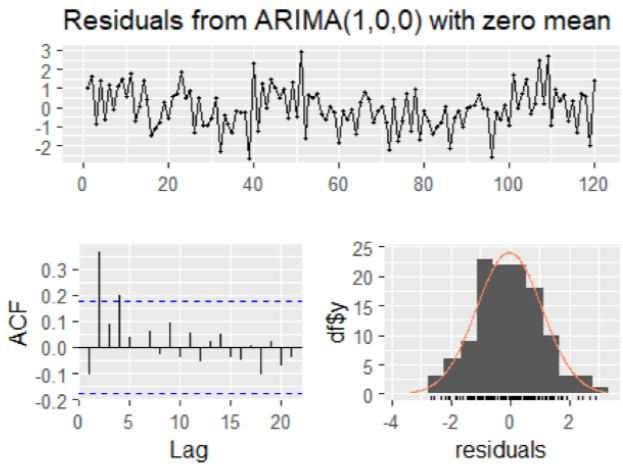
$$\hat{y}_{POP} = \frac{199.5}{6} = 33.25$$

Zadanie 5.

a) (***2p.***) Jakie cechy powinny posiadać reszty poprawnie zidentyfikowanego modelu ARMA?
b) (***1p.***) W jakim celu stosuje się test Ljunga–Boxa?
c) (***2p.***) Dla danego szeregu czasowego oszacowano trzy modele ARMA: ARMA(1,0,0), ARMA(2,0,2) oraz ARMA(2,0,1). Otrzymane wyniki zaprezentowano w tabelach 5.1, 5.2 i 5.3. Na podstawie tych danych wybierz najlepszą specyfikację modelu i uzasadnij swój wybór.

Tabela 5.1. Wyniki dla **ARIMA(1,0,0)**

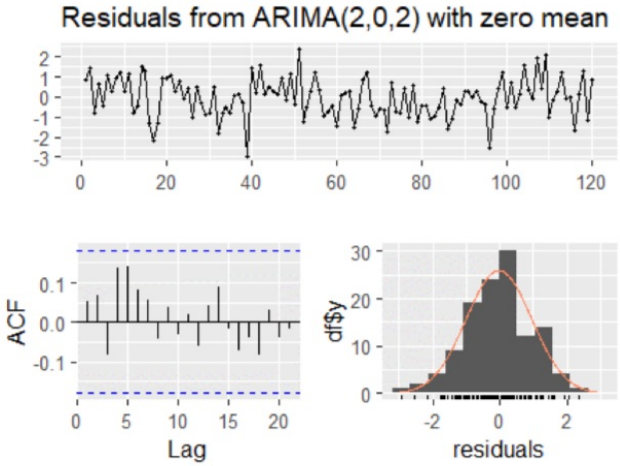
```
ARIMA(1,0,0)

Coefficients:
         ar1
      0.2527
s.e.  0.0884

sigma^2 = 1.228:  log likelihood = -
182.11
AIC=368.21   AICc=368.32   BIC=373.79
```




```
Ljung-Box test

data:  Residuals from
ARIMA(1,0,0)
Q* = 25.46, df = 9, p-value =
0.002503

Model df: 1.   Total lags
used: 10

----------------------------
Jarque - Bera Normality Test

Test Results:
  STATISTIC:
    X-squared: 0.118
P VALUE:
Asymptotic p Value: 0.9427
```

Tabela 5.2. Wyniki dla **ARIMA(2,0,2)**

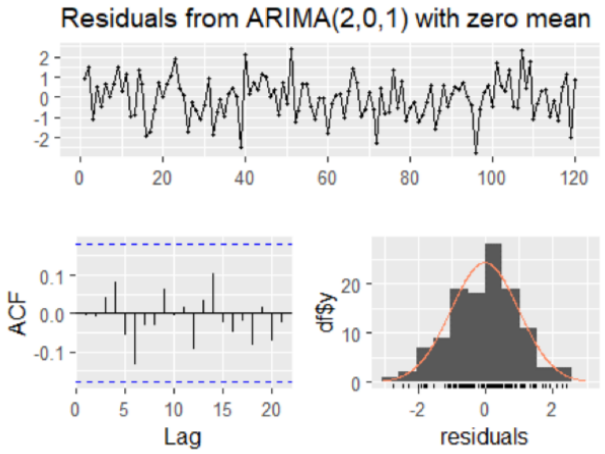
```
ARIMA(2,0,2)

Coefficients:
         ar1      ar2      ma1     ma2
      1.0213  -0.5712  -1.0049  0.9386
s.e.  0.0860   0.0968   0.0438  0.0653

sigma^2 = 0.9866:  log likelihood = -
168.75
AIC=347.49   AICc=348.02   BIC=361.43
```




```
Ljung-Box test

data:  Residuals from
ARIMA(2,0,2)
Q* = 8.4222, df = 6, p-value
= 0.2088

Model df: 4.   Total lags
used: 10

----------------------------
Jarque - Bera Normality Test

Test Results:
  STATISTIC:
    X-squared: 0.5667
P VALUE:
Asymptotic p Value: 0.7533
```

Tabela 5.3. Wyniki dla **ARIMA(2,0,1)**

```
ARIMA(2,0,1)

Coefficients:
         ar1     ar2      ma1
      0.3165  0.3761  -0.2087
s.e.  0.1789  0.1017   0.1874

sigma^2 = 1.025:  log likelihood = -
170.47
AIC=348.94   AICc=349.29   BIC=360.09
```




```
Ljung-Box test

data:  Residuals from
ARIMA(2,0,1)
Q* = 4.6212, df = 7, p-value
= 0.7061

Model df: 3.   Total lags
used: 10

----------------------------
Jarque - Bera Normality Test

Test Results:
  STATISTIC:
    X-squared: 0.4343
P VALUE:
Asymptotic p Value: 0.8048
```

a) Reszty (ang. *residuals*) poprawnie zidentyfikowanego modelu ARMA powinny posiadać cechy **białego szumu** (ang. *white noise*). Oznacza to, że reszty powinny być **nieskorelowane w czasie** i mieć **stałą wariancję**.

b) Test Ljunga-Boxa jest używany do sprawdzania, czy w szeregu czasowym występuje autokorelacja, czyli zależność między wartościami tego szeregu w różnych momentach czasowych.

c) Ljung - Box test:

H_0: Brak autokorelacji w resztach

H_1: Autokorelacja w resztach

Na podstawie tego testu należy odrzucić model ARIMA(1,0,0),

p-value = 0.0025 < 0.05 czyli reszty nie zachowują się jak biały szum.

Z pozostałych dwóch modeli ARIMA(2,0,2) ma najniższą wartość AIC czyli jest najlepiej dopasowany i powinien zostać wybrany. Można też wskazać model ARIMA(2,0,1), wartość AIC jest nieznacznie większa ale model jest prostszy (zawiera mniej zmiennych).

Zadanie 6.

Liczbę szkód w pewnym portfelu ubezpieczeń AC modelowano z uwzględnieniem dwóch następujących zmiennych objaśniających:

- *Plec* – płeć kierowcy (zmienna jakościowa: *K* (kobieta), *M* (mężczyzna)),
- *Dystans* – przebyty dystans w ciągu roku (zmienna jakościowa, przyjmująca dwie kategorie: D1 (*poniżej 20 tys. km*), D2 (*powyżej. 20 tys. km*).

Zebrano dane dotyczące liczby szkód zgłoszonych przez 3000 kierowców i przedstawiono je w tabeli 6.1 (w nawiasach podano ekspozycję na ryzyko):

Tab. 6.1

		Dystans	
		D1 (*poniżej 20 tys. km*)	D2 (*powyżej. 20 tys. km*)
Plec	*K*	15 (200)	197 (1800)
	M	28 (600)	35 (400)

Oszacowano dwa modele regresji Poissona z linkiem kanonicznym:

- M1 - w którym uwzględniono obydwie zmienne objaśniające *Plec* i *Dystans*,
- M2 – w którym uwzględniono tylko zmienną *Dystans* (pominięto *Plec*).

Uzyskano następujące oszacowania parametrów tych modeli:

Model M1:

```
Coefficients:
(Intercept)        PlecK    DystansD1
    -2.2033      -0.2853      -0.5141

Degrees of Freedom: 3 Total (i.e. Null);  1 Residual
Null Deviance:      22.76
Residual Deviance: 0.4534  AIC: 28.71
```

Model M2:

```
Coefficients:
(Intercept)    DystansD1
    -2.2495      -0.6739

Degrees of Freedom: 3 Total (i.e. Null);  2 Residual
Null Deviance:      22.76
Residual Deviance: 3.648   AIC: 29.9
```

a) (***2p.***) Przeanalizuj zależność między zmiennymi objaśniającymi (*Plec* i *Dystans*). Wiadomo, że wartość statystyki chi-kwadrat wynosi: $\chi^2 = 852.273$.

b) (***3p.***) Jak wykluczenie zmiennej silnie skorelowanej z innymi cechami wpływa na oszacowanie pozostałych współczynników regresji? Jak nazywa się to zjawisko? Wyjaśnij ten problem, odnosząc się do przykładów modeli M1 i M2.

a) Test niezależności χ^2:

H_0: Zmienne są niezależne

H_1: Zmienne nie są niezależne

Liczba stopni swobody dla tabeli kontyngencji: $(2-1)\cdot(2-1) = 1$

$\chi^2_{0.05;1} = 3.84$

Wartość statystyki testowej znajduje się w obszarze krytycznym:

$$\chi^2 = 852.273 > 3.84 = \chi^2_{0.05;1}$$

Należy odrzucić H_0 i przyjąć H_1 czyli istnieje zależność między zmiennymi.

Korzystając ze współczynnika V-Cramera:

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(k-1, r-1)}}$$

k - liczba wierszy

r - liczba kolumn

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

Wartość 0.533 wskazuje na silną zależność. Wartości bliskie 0 wskazują na niezależność.

b) Mamy tutaj do czynienia z błędem pominiętej zmiennej.
Gdy zmienne objaśniające są ze sobą skorelowane, model statystyczny ma trudność z odizolowaniem ich indywidualnego wpływu na zmienną objaśnianą. Dzieje się tak, ponieważ model, w którym brakuje jednej ze skorelowanych zmiennych, przypisuje jej brakujący wpływ pozostałym skorelowanym z nią cechom.
W przykładowych modelach widać zmianę współczynnika przy zmiennej DystansD1. Współczynnik zmienił się z -0.5141 na -0.6739 po usunięciu zmiennej PlecK.

Zadanie 7.

a) (**3*p.***) Przedstaw kluczowe etapy iteracyjnego procesu szacowania podwójnego uogólnionego modelu liniowego (DGLM - *Double Generalized Linear Model*). W jaki sposób model średniej i model dyspersji „współdziałają" podczas tego procesu?

b) (**2*p.***) Dlaczego podwójny GLM jest szczególnie istotny w modelowaniu, w którym wykorzystuje się regresję Tweedie'ego (uogólnionego modelu liniowego ze zmienną zależną o rozkładzie Tweedie z indeksem $1 < p < 2$)?

a)

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 residuów R_i^2.
3. Dopasuj GLM dla dyspersji, przyjmując jako zmienną objaśnianą kwadraty residuów 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 residuów R_i^2 i powtórz 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 obliczonie 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)

Ograniczenie standardowego modelu Tweedie GLM

Podwójny uogólniony model liniowy (DGLM) jest szczególnie istotny w modelowaniu z wykorzystaniem regresji Tweedie'ego, ponieważ standardowy model Tweedie GLM narzuca **nierealistyczne ograniczenie** dotyczące wpływu czynników ryzyka na częstość i wysokość szkód.

W standardowym modelu GLM parametr dyspersji ϕ musi być **stały** dla wszystkich obserwacji. W przypadku rozkładu Tweedie, który jest modelem złożonym Poissona-Gamma, parametr dyspersji ϕ zależy zarówno od parametrów rozkładu częstości szkód (Poisson), jak i rozkładu wysokości pojedynczej szkody (Gamma).

To założenie o stałości ϕ prowadzi do następującego, kluczowego ograniczenia:

> Każdy czynnik ryzyka, który zwiększa oczekiwaną wysokość szkody, musi również zwiększać oczekiwaną częstość szkód (i odwrotnie). Wpływ danego czynnika na częstość i wysokość szkody **musi iść w tym samym kierunku**.

Podwójny GLM (DGLM) rozwiązuje ten problem, ponieważ **pozwala, aby parametr dyspersji ϕ również zależał od zmiennych objaśniających**. Modelując dyspersję za pomocą osobnego zestawu predyktorów, DGLM uwalnia sztywną zależność między komponentami częstości i wysokości szkody. Dzięki temu model może prawidłowo odzwierciedlić sytuacje, w których dany czynnik ryzyka ma różny lub nawet przeciwstawny wpływ na częstość i wysokość szkód, co prowadzi do znacznie bardziej wiarygodnych wyników.

Zadanie 8.

W pewnym portfelu ubezpieczeń przez rok (365 dni) rejestrowano dzienną liczbę szkód. Uzyskano następujące wyniki (tab. 8.1):

Tab. 8.1

Liczba szkód	Liczba dni
0	50
1	122
2	101
3	92
≥ 4	0

a) (***1p.***) Przyjmij rozkład Poissona dla dziennej liczby szkód i oszacuj jego parametr.

b) (***4p.***) Sprawdź dopasowanie tego rozkładu testem zgodności chi-kwadrat. Przyjmij poziom istotności 0.05.

a) Estymator największej wiarygodności dla parametru λ rozkładu Poissona jest równy średniej arytmetycznej z próby, $\bar{X}$.

Liczba szkód k_i	Liczba dni O_i	$O_i k_i$
0	50	0
1	122	122
2	101	202
3	92	276
Σ	**365**	**600**

$$\hat{\lambda} = \bar{x} = \frac{600}{365} = 1.6438$$

b) Test zgodności χ^2:

H_0: dane pochodzą z rozkładu Poissona z parametrem $\lambda = 1.6438$

H_1: dane nie pochodzą z powyższego rozkładu

Masa p-stwa rozkładu Poissona: $p_k = \frac{e^{-\hat{\lambda}} \hat{\lambda}^k}{k!}$

Statystyka testowa: $\chi^2 = \sum_{i=1}^{m} \frac{(E_i - O_i)^2}{E_i}$

Liczba szkód k_i	Liczba dni O_i	E_i	$\frac{(E_i - O_i)^2}{E_i}$
0	50	$365 \cdot e^{-1.6438} = 70.534$	5,978
1	122	$365 \cdot 1.6438 \cdot e^{-1.6438} = 115.944$	0,316
2	101	$365 \cdot \frac{1.6438^2}{2} \cdot e^{-1.6438} = 95.294$	0,342
≥ 3	92	$365 - (70.534 + 115.944 + 95.294) = 83.227$	0,925
		Σ	**7,561**

$\chi^2 = 7.561$

Liczba stopni swobody : liczba grup - liczba parametrów - 1 = 4 - 1 - 1 = 2

Wartość krytyczna :

$\chi^2_{0.05;2} = 5.991$

$\chi^2 = 7.561 > 5.991 = \chi^2_{0.05;2}$

Należy odrzucić H_0 i przyjąć H_1 czyli dane nie pochodzą z rozkładu Poissona z parametrem $\lambda = 1.6438$.

Zadanie 9.

Oszacowano model klasyfikacyjny do przewidywania wystąpienia oszustwa w zgłaszanych roszczeniach ubezpieczeniowych. Model ten został oceniony na zbiorze testowym zawierającym 10 obserwacji. W tabeli 9.1 przedstawiono rzeczywiste wartości zmiennej objaśnianej y_i (1 – oszustwo, 0 – brak oszustwa) oraz prognozowane prawdopodobieństwa $\hat{p}_i$ wystąpienia oszustwa, uzyskane z tego modelu.

Tabela 9.1

y_i	1	1	0	0	1	0	0	1	1	0
$\hat{p}_i$	0.40	0.20	0.22	0.73	0.90	0.34	0.20	0.80	0.45	0.34

a) (***2p.***) Narysuj krzywą ROC dla wartości progowej (punktu odcięcia) 0.40. Opisz osie wykresu.

b) (***1p.***) Narysuj krzywe ROC dla
   i. modelu nie posiadającego żadnych zdolności predykcyjnych;
   ii. hipotetycznego „idealnego" modelu.

c) (***2p.***) Jak zmiana progu wpływa na wynikowy zysk lub straty finansowe firmy, biorąc pod uwagę koszty niewykrycia oszustwa oraz koszty błędnej identyfikacji niewinnego roszczenia jako oszukańczego?

a) Należy obliczyć czułość i 1- specyficzność.

$$\text{Czułość} = \frac{\text{true positive}}{\text{all positive}} = \frac{TP}{P}$$

$$\text{Specyficzność} = \frac{\text{true negative}}{\text{all negative}} = \frac{TN}{N}$$

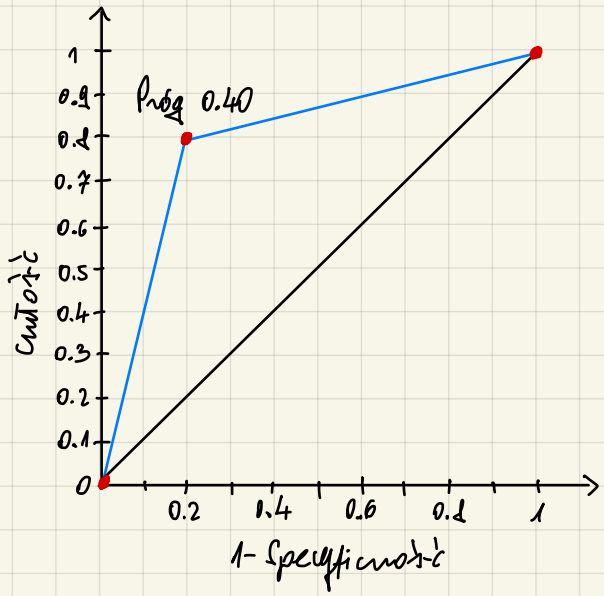
Dla progu równego 0.40, każdą obserwację z przewidywanym p-stwem $\hat{p}_i \geq 0.40$ klasyfikujemy jako oszustwo (predykcja = 1).

Tabela trafności dla progu 0.40

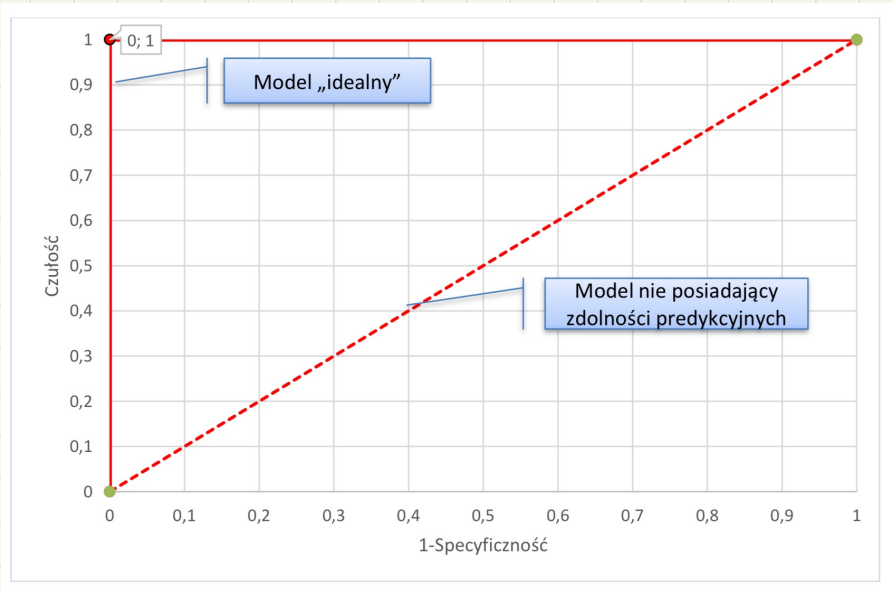
Prognozy	Faktyczne		Razem
	N	P	
N	4	1	5
P	1	4	5
Razem	5	5	10

$$\text{Czułość} = \frac{4}{5} = 0.8$$

$$1 - \text{Specyficzność} = 1 - \frac{4}{5} = 0.2$$



b)



c)

Podniesienie progu

Podwyższenie progu sprawia, że model zmniejsza liczbę przypadków oznaczonych jako oszukańcze. Oznacza to, że więcej rzeczywistych oszustw pozostaje niewykrytych. Firma traci pieniądze na wypłatach, które nie powinny mieć miejsca. Koszt niewykrycia jednego oszustwa może być wysoki, zwłaszcza gdy wartość roszczenia jest duża. W przypadku bardzo wysokich kosztów niewykrycia oszustwa wzrost progu może prowadzić do znacznych strat, szczególnie jeśli oznacza to liczne, niewykryte przypadki oszustwa.

Obniżenie progu

Obniżenie progu skutkuje większą liczbą przypadków oznaczonych jako oszustwa, nawet gdy są to faktycznie roszczenia prawdziwe. Wiąże się to z tym, że więcej niewinnych roszczeń zostaje błędnie sklasyfikowanych jako oszukańcze. Może to generować koszty administracyjne, ponieważ każda podejrzana sprawa wymaga dodatkowego procesu weryfikacji. Dodatkowo odrzucone niesłusznie roszczenia mogą prowadzić do niezadowolenia klientów, zwiększenia liczby reklamacji i potencjalnych strat związanych z utratą klientów. Jeżeli zatem koszty dodatkowej weryfikacji i utraty klientów przeważają nad korzyściami wykrycia dodatkowych oszustw, obniżenie progu może zmniejszać zyski firmy.

Zadanie 10.

a) (***2p.***) W jaki sposób analiza składowych głównych (PCA, *Principal Component Analysis*) może być wykorzystana do analizy danych ubezpieczeniowych, takich jak ryzyko klienta lub analiza szkód?

b) (***1p.***) Dlaczego PCA jest uważana za metodę uczenia bez nadzoru (*unsupervised*)?

c) (***2p.***) Firma ubezpieczeniowa analizuje ryzyko klientów na podstawie trzech zmiennych:

- Wiek (X1)
- Dochód roczny (X2)
- Liczba zgłoszonych szkód (X3)

Macierz kowariancji Σ dla tych zmiennych wygląda następująco:

$$\Sigma = \begin{bmatrix} 2.0 & 0.8 & 0.4 \\ 0.8 & 1.0 & 0.7 \\ 0.4 & 0.7 & 1.0 \end{bmatrix}$$

Wartości własne tej macierzy są równe (po zaokrągleniu do trzech miejsc po przecinku):

$$\lambda_1 = 2.731, \quad \lambda_2 = 1.032, \quad \lambda_3 = ?\,.$$

- Oblicz, jaką część wyjściowej wariancji wyjaśnia każda z trzech głównych składowych.
- Zdecyduj, ile składowych należy zachować, aby wyjaśnić co najmniej 90% wariancji danych.

a) Na przykład do:

- redukcji zmiennych wpływających na profil ryzyka,
- segmentacji klientów według ryzyka,
- identyfikacji klientów o nietypowych profilach ryzyka,
- identyfikacji kluczowych czynników wpływających na wielkość szkód,
- monitorowania zmian w szkodowości portfela.

b) W metodach uczenia bez nadzoru model jest trenowany wyłącznie na podstawie danych wejściowych bez wiedzy o konkretnym wyniku czy klasie, do której miałby przyporządkować obserwacje. PCA analizuje wyłącznie struktury i zależności między zmiennymi w danych wejściowych, identyfikując ich korelacje, a następnie przekształca zbiór danych tak, aby uzyskać nowy zestaw zmiennych (składowe główne). W praktyce oznacza to, że PCA szuka wzorców i ukrytych struktur w danych, co czyni ją idealnym narzędziem do eksploracyjnej analizy danych, segmentacji czy redukcji wymiarowości,bez konieczności przypisania punktów danych do określonych kategorii czy przewidywania wyników.

c) Suma wartości własnych macierzy jest równa jej śladowi, czyli sumie elementów na głównej przekątnej, stąd:

$\lambda_3 = 2+1+1 - 2.731 - 1.032 = 0.237$

Aby obliczyć jaką część wariancji wyjaśnia każda ze składowych głównych, dzielimy jej wartość własną przez całkowitą wariancję.

Wariancja wyjaśniona przez:

- pierwszą składową : $\frac{1}{4} \cdot 2.731 \cdot 100\% = 68.27\%$
- drugą składową : $\frac{1}{4} \cdot 1.032 \cdot 100\% = 25.80\%$
- trzecią składową : $\frac{1}{4} \cdot 0.237 \cdot 100\% = 5.93\%$

Należy wybrać dwie pierwsze składowe, tłumaczą 94.07 % wariancji.