Zadanie 1.

Dla pewnego portfela ubezpieczeń badano zależność rocznej liczby szkód (zmienna *clm.count*) od wieku ubezpieczonego wyrażonego w latach (zmienna ilościowa *driver.age*) oraz płci (zmienna jakościowa *driver.gender*, przyjmująca dwie wartości: *Female*, *Male*). Oszacowano dwa modele regresji Poissona z kanonicznymi funkcjami łączącymi (linkami kanonicznymi). W obydwu modelach jako zmienną offsetową uwzględniono czas ekspozycji na ryzyko w latach (zmienna *exposure*). Uzyskano następujące wyniki:

Model M1:

```
Call:
glm(formula = clm.count ~ driver.age + driver.gender + offset(log(exposure)),
   family = poisson, data = zbior.uczacy)
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.6960  -0.4767  -0.3828  -0.2566   4.8785
Coefficients:
                      Estimate      Std. Error     z value       Pr(>|z|)
(Intercept)           -1.231453     0.107225       -11.485       < 2e-16 ***
driver.age            -0.009330     0.002082       -4.482        7.39e-06 ***
driver.genderMale     -0.189300     0.065907       -2.872        0.00408 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)
Null deviance: 9881.4  on 24455  degrees of freedom
Residual deviance: 9852.1  on 24453  degrees of freedom
'log Lik.' -6865.075
```

Model M2 (zmienna $driver.age.kw = driver.age^2$)**:**

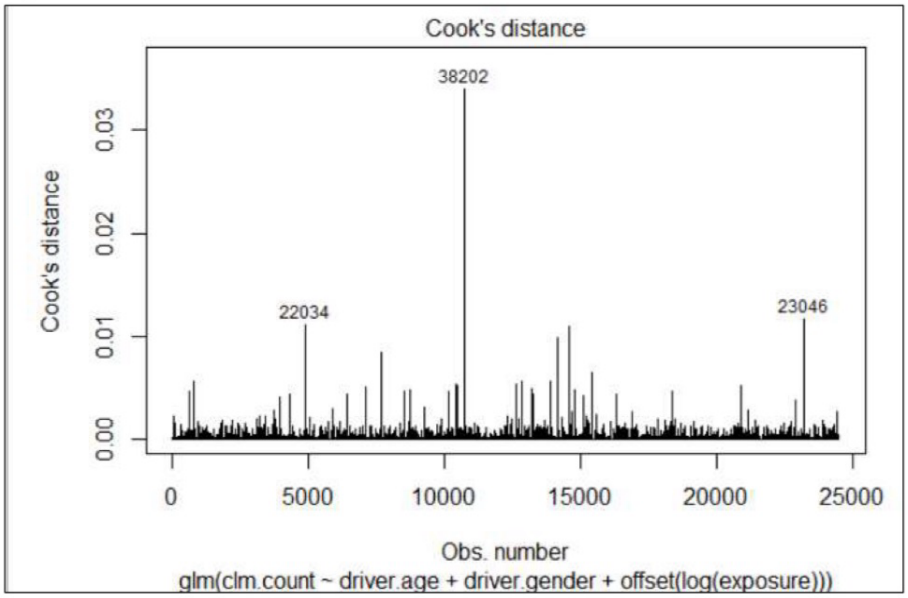
```
Call:
glm(formula = clm.count ~ driver.age + driver.age.kw + driver.gender +
   offset(log(exposure)), family = poisson, data = zbior.uczacy)
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.9647  -0.4706  -0.3824  -0.2692   4.8862
Coefficients:
                      Estimate      Std. Error     z value       Pr(>|z|)
(Intercept)           -3.647114     0.530283       -6.878        6.08e-12 ***
driver.age             0.018353     0.006276       2.924         0.00345 **
driver.age.kw          0.000493     0.000106       4.665         3.09e-06 ***
driver.genderMale     -0.187095     0.065926       -2.838        0.00454 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)
Null deviance: 9881.4  on 24455  degrees of freedom
Residual deviance: 9831.9  on 24452  degrees of freedom
'log Lik.' -6855
```

a) (***2p.***) Wyjaśnij związek między wiarygodnością *L,* a kryterium informacyjnym AIC i wskaż kiedy każdy z tych mierników może być użyty do porównania różnych modeli. Który z oszacowanych modeli (tzn. M1, M2) jest lepszy? Wybór uzasadnij.

b) (***2p.***) W zbiorze uczącym wykorzystanym do oszacowania obydwu modeli znajduje się 46-ścio letnia kobieta z trzymiesięczną ekspozycją na ryzyko, w czasie której nie zgłosiła żadnej szkody. Wykorzystując model M1:
   - oszacuj prawdopodobieństwo, że w ciągu jednego roku nie zgłosi ona żadnej szkody,
   - wyznacz resztę Pearsona odpowiadającą tej obserwacji.

c) (***1p.***) Na rysunku 1.1 przedstawiono jeden z wykresów diagnostycznych dla modelu M1. Wyjaśnij w jakim celu wykorzystuje się tego typu wykresy. Czy dla modelu M2 uzyskamy identyczny? Odpowiedź uzasadnij.

Rys. 1.1



a) $AIC = -2\ln(L) + 2p$, gdzie $\ln(L)$ - logarytm funkcji wiarygodności

p - liczba obserwowanych parametrów

Wiarygodność można wykorzystać do porównywania modeli które posiadają taką samą liczbę parametrów lub są zagnieżdżone.

AIC jest bardziej przydatne, gdy modele różnią się liczbą parametrów i są zbudowane na innym zestawie zmiennych objaśniających.

$AIC_{M1} = 2 \cdot 3 - 2 \cdot (-6865.075) = 13736.15$

$AIC_{M2} = 2 \cdot 4 - 2 \cdot (-6855) = 13718$

$AIC_{M2} < AIC_{M1}$ - model M2 jest lepszy

b) Ekspozycja jest równa więc wynosi 1:

$\hat{\mu} = \exp\{-1.231453 - 0.00933 \cdot 46 - 0.1293 \cdot 0\} = 0.190019$

Rozkład Poissona:

$P(Y=k) = \frac{1}{k!} e^{-\hat{\mu}} \hat{\mu}^{k}$

$P(Y=0) = \exp\{-0.190019\} = 0.8269433$

Reszta Pearsona w modelu Poissona:

$r_p = \frac{y_i - \mu_i}{\sqrt{\mu_i}}$

Ekspozycja dla tej obserwacji wynosi 3 miesiące czyli 0.25 :

$$r_p = \frac{0 - 0.25 \cdot 0.190019}{\sqrt{0.25 \cdot 0.190019}} = -0.2179559$$

c) Wykres „Cook's distance" jest wykorzystywany w analizie regresji jako miara wpływu poszczególnych obserwacji na wynik regresji. Umożliwia wykrycie obserwacji, które znacząco wpływają na wynik regresji, a tym samym pozwala zbadać ich wpływ na model.

Dla M2 wyszła inny wykres, ponieważ w mierze Cooka uwzględnia się reszty modeli.

Zadanie 2.

a) **(1*p*.)** Wyjaśnij w jaki sposób przeprowadza się *k*-krotną walidację krzyżową.

b) **(1*p*.)** Podaj na czym polega walidacja za pomocą metody LOOCV (*Leave-one-out cross-validation*).

c) **(2*p*.)** Jakie są zalety i wady *k*-krotnej walidacji krzyżowej w porównaniu z:

   i. podejściem wykorzystującym jedynie jeden zbiór walidacyjny,

   ii. metodą LOOCV.

   W odpowiedzi uwzględnij problem kompromisu między obciążeniem a wariancją modelu.

d) **(1*p*.)** Oszacowano model regresji liniowej na podstawie 5-ciu obserwacji. Uzyskano następujące reszty: 1.78, -1.30, 1.09, -1.89, 0.32. Wiadomo, że w analizowanym przypadku macierz daszkowa jest równa:

$$H = \begin{bmatrix} 0.29 & 0.24 & -0.03 & 0.15 & 0.35 \\ & 0.22 & 0.09 & 0.17 & 0.27 \\ & & 0.80 & 0.34 & -0.20 \\ & & & 0.23 & 0.11 \\ & & & & 0.46 \end{bmatrix}$$

Walidację tego modelu przeprowadzono z wykorzystaniem błędu średniokwadratowego MSE (*mean squared error*) za pomocą metody LOOCV. Jaki otrzymano wynik?

a)

1. **Podział danych**: Dostępny zbiór obserwacji jest losowo dzielony na *k* grup (nazywanych też *zbiorami* lub *foldami*) o w przybliżeniu równej wielkości.

2. **Iteracyjne trenowanie i walidacja**:
   - Pierwsza grupa (zbiór 1) jest traktowana jako **zbiór walidacyjny**, a model jest trenowany na pozostałych *k-1* grupach (które łącznie tworzą zbiór uczący).
   - Oblicza się błąd predykcji (np. błąd średniokwadratowy, MSE) dla obserwacji w zbiorze walidacyjnym.
   - Proces ten jest powtarzany *k* razy, przy czym za każdym razem inna grupa pełni rolę zbioru walidacyjnego.

3. **Obliczenie ostatecznego wyniku**: Po przeprowadzeniu *k* iteracji uzyskuje się *k* różnych estymacji błędu testowego ($MSE_1, MSE_2, ..., MSE_k$). Ostateczną estymacją błędu w metodzie k-krotnej walidacji krzyżowej jest średnia z tych wartości:

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

b) Walidacja krzyżowa z pominięciem jednej obserwacji (Leave-one-out cross-validation, LOOCV) to intensywna obliczeniowo, ale czasami użyteczna metoda szacowania błędu testowego modelu statystycznego. Jest to szczególny przypadek k-krotnej walidacji krzyżowej, w którym liczba podzbiorów (k) jest równa liczbie obserwacji (n) w zbiorze danych.

Proces polega na wielokrotnym dopasowywaniu modelu, gdzie za każdym razem jedna obserwacja jest wykluczana ze zbioru uczącego i używana do walidacji.

c) Wybór *k* w walidacji krzyżowej to **kompromis między obciążeniem a wariancją**.

- **LOOCV (k=n)** ma niskie obciążenie, ale wysoką wariancję.
- **Podejście z jednym zbiorem walidacyjnym (odpowiednik k=2)** ma wysokie obciążenie, ale niską wariancję.

K-krotna walidacja krzyżowa z wartościami **k=5 lub k=10** jest w praktyce złotym środkiem. Empirycznie wykazano, że takie wartości prowadzą do estymacji błędu testowego, które nie cechują się ani nadmiernie wysokim obciążeniem, ani bardzo wysoką wariancją.

d)

$$CV_n = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1-h_i}\right)^2$$

$$CV_n = \frac{1}{5}\left[\left(\frac{1.71}{1-0.29}\right)^2 + \left(\frac{-1.30}{1-0.22}\right)^2 + \left(\frac{1.09}{1-0.2}\right)^2 + \left(\frac{-1.29}{1-0.23}\right)^2 + \left(\frac{0.32}{1-0.46}\right)^2\right] =$$

$$= 9.028$$

Zadanie 3.

a) (**2*p.***) Krótko przedstaw ideę uogólnionych modeli addytywnych (*Generalized Additive Models* – GAM). Wskaż dlaczego weszły do zestawu narzędzi aktuariusza.

b) (**1*p.***) Podaj definicję funkcji sklejanej stopnia 3 (splajnu kubicznego).

c) (**2*p.***) Liczbę roszczeń (zmienna *clm.count*) w pewnym portfelu ubezpieczeń AC modelowano z uwzględnieniem następujących zmiennych objaśniających:

driver.gender – płeć kierowcy (zmienna jakościowa: *Female*, *Male*),
driver.age – wiek kierowcy (zmienna ilościowa),
vehicle.age - wiek samochodu (zmienna ilościowa),
vehicle.value – wartość samochodu (zmienna ilościowa),
hp – moc silnika (zmienna ilościowa).

Oszacowano uogólniony model addytywny, w którym przyjęto rozkład Poissona dla liczby roszczeń oraz link logarytmiczny. Zinterpretuj uzyskane wyniki (podane poniżej). W interpretacji uwzględnij także wykresy przedstawione na rysunku 3.1.

Family: poisson
Link function: log

Formula:
clm.count ~ driver.gender + s(driver.age) + s(vehicle.age) + s(vehicle.value, hp) + offset(exposure)

Parametric coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.86959	0.07887	-36.382	< 2e-16 ***
driver.genderMale	-0.23469	0.08422	-2.787	0.00533 **

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

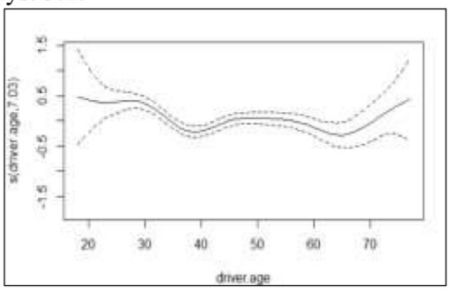
Approximate significance of smooth terms:

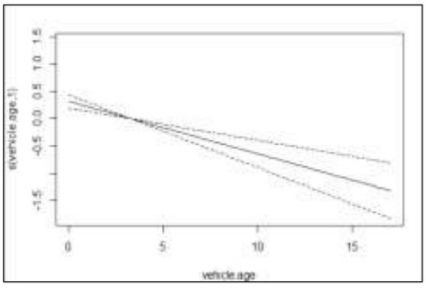
	edf	Ref.df	Chi.sq	p-value
s(driver.age)	7.026	8.014	43.38	< 2e-16 ***
s(vehicle.age)	1.001	1.002	26.86	2.41e-07 ***
s(vehicle.value,hp)	3.320	4.299	17.34	0.00224 **

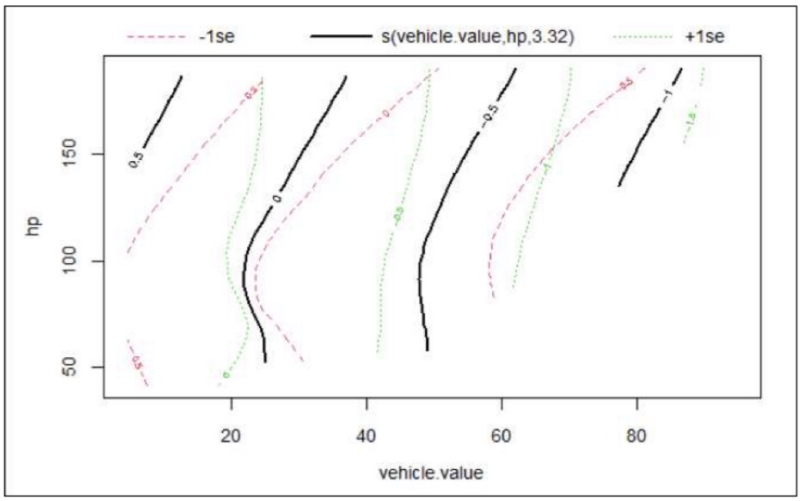
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) = 0.0264 Deviance explained = 1.66%
UBRE = -0.59697 Scale est. = 1 n = 14634

Rys. 3.1.







a) Uogólnione modele addytywne (GAM) to rozszerzenie uogólnionych modeli liniowych (GLM), które pozwala na modelowanie nieliniowych zależności między predyktorami a zmienną odpowiedzi, zachowując przy tym addytywną strukturę. Zamiast zakładać, że każdy predyktor ma liniowy wpływ na odpowiedź (np. $\beta_1 x_1$), GAM dopasowuje gładką, nieliniową funkcję dla każdego predyktora (np. $f_1(x_1)$). Model końcowy jest sumą tych indywidualnych funkcji: $Y = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p) + \epsilon$. Taka budowa pozwala na dużą elastyczność w modelowaniu złożonych wzorców, jednocześnie zachowując możliwość interpretacji wpływu każdego predyktora z osobna.

Dzięki GAM aktuariusze mogą:

- Precyzyjnie modelować **nieliniowe zależności** i **interakcje** między zmiennymi, takimi jak wiek i płeć kierowcy czy dane geograficzne.
- Analizować indywidualne dane dotyczące śmiertelności przy użyciu **regresji Poissona** bez konieczności wstępnego, subiektywnego grupowania danych.

b) Splajn sześcienny z K węzłami można zamodelować jako:

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+3} b_{K+3}(x_i) + \epsilon_i,$$

dla odpowiedniego wyboru funkcji bazowych $b_1, b_2, \ldots, b_{K+3}$. Najbardziej bezpośrednim sposobem reprezentacji splajnu sześciennego przy użyciu powyszej jest rozpoczęcie od bazy dla wielomianu sześciennego – mianowicie, $b_1(x_i) = x_i$, $b_2(x_i) = x_i^2$, i $b_3(x_i) = x_i^3$ a następnie dodanie jednej **potęgowej funkcji bazowej** na każdy węzeł. Potęgowa funkcja bazowa jest zdefiniowana jako:

$$h(x_i, \xi_i) = (x_i - \xi_i)_+^3 = \begin{cases} (x_i - \xi_i)^3 & \text{jeśli } x_i > \xi_i \\ 0 & \text{w przeciwnym razie,} \end{cases}$$

gdzie ξ_i jest i-tym węzłem. Przykład:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - \xi)_+^3$$

c) Efekty liniowe:

- driver.gender: poziomem odniesienia jest Female. Współczynnik dla Male wynosi −0.23469 i jest istotny statystycznie p-value < 0.05.

Efekty nieliniowe:

- driver.age: edf = 7.026 => silny nieliniowy wpływ na wynik co potwierdza wykres, krzywa ma kształt wielomianowy. p-value < 0.05 więc zmienna jest istotna.
- vehicle.age: edf = 1.001 => wpływ raczej liniowy co potwierdza wykres,

linia prosta. p-value < 0.05 więc zmienna jest istotna.

- Interakcja (vehicle.value, hp): edf = 3.32 => wpływ nieliniowy, p-value < 0.05 więc zmienna istotna. Wykres konturowy:
  - dla pojazdów o niskiej wartości i wysokiej mocy widać dodatni wpływ na wysyłkę
  - dla pojazdów o wysokiej wartości i wysokiej mocy widać ujemny wpływ na wysyłkę

Zadanie 4.

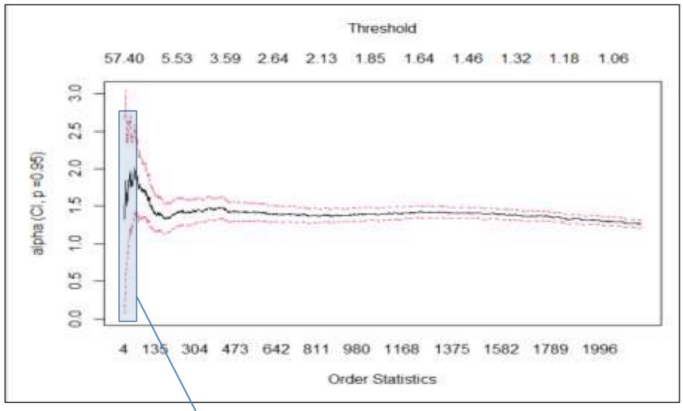
a) (**2*p.***) Krótko opisz podejście Hilla do modelowania ogonów rozkładów (m in. podaj założenia odnośnie rozkładów i przedstaw odpowiedni estymator).

b) (**1*p.***) Przedstaw konstrukcję wykresu Hilla (*Hill plot*) i wskaż w jakim celu jest wykorzystywany.

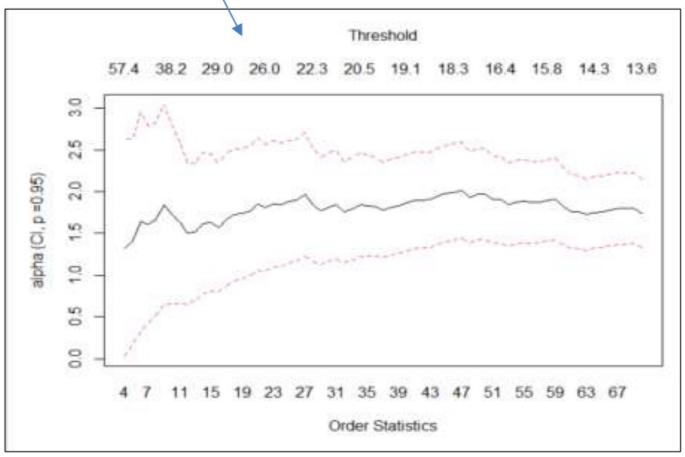
c) (**2*p.***) Analizowano straty pożarowe zarejestrowane przez Copenhagen Re. Na poniższym rysunku (Rys. 4.1.) przedstawiono skonstruowany na ich podstawie wykres Hilla (panel **B** przedstawia powiększony fragment zaznaczony na panelu **A**). Zinterpretuj otrzymane wyniki.

Rys. 4.1.

A



B



a)

Metoda Hilla jest podejściem stosowanym do modelowania i estymacji ogonów rozkładów prawdopodobieństwa, zwłaszcza tych o grubych ogonach (heavy-tailed). Służy do oszacowania parametru kształtu ogona, znanego jako indeks ogona.

Podstawowym założeniem metody Hilla jest to, że funkcja przeżycia (ogon) badanego rozkładu dla dużych wartości x zachowuje się jak funkcja potęgowa:

$$\bar{F}(x) = x^{-\alpha} L(x)$$

gdzie:

- $\alpha > 0$ to **indeks ogona**, który opisuje "ciężkość" ogona (im mniejsze α, tym grubszy ogon).
- $L(x)$ jest funkcją **wolno zmienną**, co oznacza, że zmienia się wolniej niż jakakolwiek funkcja potęgowa.

Estymator Hilla:

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

gdzie:

- k to liczba największych obserwacji użytych do estymacji ($2 \leq k \leq n$).
- $X_{j,n}$ to j-ta największa obserwacja w próbie.

b) Aby wybrać optymalną wartość k, tworzy się tzw. **wykres Hilla**, który przedstawia wartości estymatora $\hat{\alpha}_{k,n}^{(H)}$ w funkcji k. Następnie poszukuje się na wykresie stabilnego regionu, w którym estymaty są względnie stałe, i na tej podstawie wybiera się ostateczne oszacowanie indeksu ogona α.

Aby k-ty moment statystyczny (jak średnia czy wariancja) był skończony, wartość indeksu α musi być od tego k większa. Np. wariancja (związana z 2 momentem, $k = 2$) jest skończona, jeśli $\alpha > 2$.

c) Przedstawione wykresy można wykorzystać do oceny indeksu ogona rozkładu α. W przypadku analizowanych danych można przyjąć, że α wynosi od 1.5 do 2 co sugeruje, że dane pochodzą z rozkładu o grubym ogonie. Ponieważ $\alpha \leq 2$ to wariancja tego rozkładu jest nieskończona.

Zadanie 5.

a) (**2*p.***) Wymień etapy statystycznej analizy szeregów czasowych danych $y_1, y_2, \dots, y_t$. Krótko opisz jeden z nich.

b) (**2*p.***) Przedstaw sposób prognozowania szeregów czasowych za pomocą modeli ARMA. Podaj ogólne założenia i wskaż ideę.

c) (**1*p.***) Na podstawie szeregu czasowego liczącego 200 obserwacji oszacowano model ARMA(1,1). Uzyskano następujące wyniki:

```
Call:
arima(x = data, order = c(1, 0, 1), method = "ML")
Coefficients:
      ar1     ma1     intercept
   0.4039  0.5361    0.0393
s.e. 0.0788  0.0668    0.1866

sigma^2 estimated as 1.059:  log likelihood = -290,  aic = 588
```

Wartości rzeczywiste i oszacowane reszty $\hat{\varepsilon}_t$ dla 3 ostatnich obserwacji przedstawia tabela 5.1.

Tab. 5.1

t	*198*	*199*	*200*
x_t	1.17510868	-0.11635671	0.06456704
$\hat{\varepsilon}_t$	-0.6482727	-0.2668962	0.2312339

Wyznacz prognozę dla tego szeregu czasowego na okres $t = 202$.

a)

Etapy statystycznej analizy szeregów czasowych:

- Analiza wstępna
- Analiza w dziedzinie czasu
- Dopasowanie modelu
- Analiza reszt i porównanie modeli

Analiza wstępna:

Jest to pierwszy etap, w którym dane są wizualizowane na wykresie w celu oceny, czy zastosowanie pojedynczego modelu stacjonarnego jest zasadne. Analiza wstępna obejmuje również rozważenie, czy konieczna jest wstępna obróbka danych, na przykład w celu usunięcia trendów lub sezonowości. Istotnym elementem tego etapu jest wybór odpowiedniej długości okna czasowego dla danych, co wiąże się z kompromisem między potrzebą korzystania z najbardziej aktualnych informacji a koniecznością posiadania wystarczająco dużej próbki do precyzyjnej estymacji statystycznej.

b)

Podstawowym założeniem jest to, że dane pochodzą z odwracalnego modelu ARMA, a innowacje (błędy) procesu mają własność różnicy martyngałowej. Oznacza to, że oczekiwana wartość przyszłej innowacji, biorąc pod uwagę historię procesu, jest równa zero.

Główną ideą jest wykorzystanie warunkowej wartości oczekiwanej $E(X_{t+h} \mid \mathcal{F}_t)$ jako predyktora, gdzie $\mathcal{F}_t$ reprezentuje historię procesu do czasu t. Ten predyktor minimalizuje średniokwadratowy błąd prognozy. Prognozy oblicza się rekurencyjnie. Wartości losowe do czasu t są traktowane jako "znane", a oczekiwane wartości przyszłych innowacji (dla $h \geq 1$) wynoszą zero. W praktyce, ponieważ pełna historia procesu nie jest znana, do obliczeń wykorzystuje się reszty z dopasowanego modelu. W miarę wydłużania horyzontu prognozy, przewidywana wartość zbiega do bezwarunkowej średniej procesu.

c) Predykcja w modelu ARMA(p,q):

$$X_t = \mu + \sum_{i=1}^{p} \varphi_i (X_{t-i} - \mu) + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}$$

ARMA(1,1):

$$X_t = \mu + \varphi_1 (X_{t-1} - \mu) + \theta_1 \epsilon_{t-1}$$

$$x^p_{201} = 0.0393 + 0.4039(0.06456704 - 0.0393) + 0.5361 \cdot 0.2312339 =$$

$$= 0.1734699$$

$\epsilon_{201} = 0$ wartość przyszłego błędu jest niemana więc jej wartość oczekiwana wynosi 0.

$$x^p_{202} = 0.0393 + 0.4039(0.1734699 - 0.0393) = 0.0934912$$

Zadanie 6.

a) (**2p.**) Podaj definicję danych prawostronnie cenzurowanych (*right censoring*). Wskaż i omów co najmniej dwie sytuacje, w których aktuariusz analizuje tego typu dane.

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

i. (**2p.**) Skonstruuj estymator Kaplana–Meiera dla funkcji przeżycia $S(x)$.

ii. (**1p.**) Oszacuj wariancję estymatora Kaplana–Meiera dla $S(2)$.

Tab. 6.1

1	2	3*	4	4	4*	4*	5	7*	8	8	8	9	9	9	9	10*	12	12	15*

a) Obserwacja cenzurowana prawostronnie w punkcie u, jeśli gdy jest równa lub większa od u, jest zapisywana jako równa u, ale gdy jest poniżej u, jest zapisywana z jej obserwowaną wartością.

W danych dotyczących roszczeń ubezpieczeniowych, obecność limitu polisy może prowadzić do prawostronnie cenzurowanych obserwacji. Gdy kwota szkody jest równa lub przekracza limit u, świadczenia powyżej tej wartości nie są wypłacane, więc dokładna wartość zazwyczaj nie jest rejestrowana. Wiadomo jednak, że wystąpiła szkoda o wartości co najmniej u.

Podczas przeprowadzania badania śmiertelności ludzi, jeśli osoba żyje w momencie zakończenia badania, nastąpiło cenzurowanie prawostronne. Wiek osoby w chwili śmierci nie jest znany, ale wiadomo, że jest on co najmniej tak duży jak wiek w momencie zakończenia badania.

b) Estymator Kaplana - Meiera:

$$\hat{S}_n(y) = \prod_{i: y_i \le y} \left(1 - \frac{d_i}{r_i}\right)$$

i	y_i	d_i	b_i	r_i	$\hat{S}_n(y_i)$
1	1	1	0	20	$1 - \frac{1}{20} = 0.950$
2	2	1	1	19	$0.950(1 - \frac{1}{19}) = 0.900$
3	4	2	2	17	$0.900(1 - \frac{2}{17}) = 0.749$
4	5	1	1	13	$0.749(1 - \frac{1}{13}) = 0.733$
5	8	3	0	11	$0.733(1 - \frac{3}{11}) = 0.533$
6	9	4	1	8	$0.533(1 - \frac{4}{8}) = 0.267$
7	12	2	1	3	$0.267(1 - \frac{2}{3}) = 0.089$

Wzór Greenwooda:

$$Var[S_n(y)] = [S_n(y)]^2 \sum_{i: y_i \le y} \frac{d_i}{n_i(n_i - d_i)}$$

$$Var[S_{20}(2)] = 0.9^2 \left[\frac{1}{20 \cdot 19} + \frac{1}{19 \cdot 18} \right] = 0.0045$$

Zadanie 7.
Przedstaw wytyczne Krajowego Standardu Aktuarialnego w zakresie stosowania modeli (tj. wyboru, tworzenia, modyfikowania i przeliczania modeli) dotyczące:

a) (**1*p.***) ryzyka modelu,
b) (**2*p.***) walidacji modeli,
c) (**2*p.***) wykorzystania wyników przebiegu modelu.

a) Aktuariusz powinien zapewnić, że ryzyka modelu zostały zidentyfikowane, ocenione i że istnieją odpowiednie działania mające na celu ograniczenie takich ryzyk, takie jak odpowiednia walidacja modelu, dokumentacja i kontrola.

b) Aktuariusz powinien mieć pewność, że została przeprowadzona odpowiednia walidacja modelu. Walidacja modelu obejmuje ocenę, czy:

- Model jest dopasowany do zamierzonego celu prac. Kwestie, które aktuariusz powinien rozważyć, tam gdzie mają one zastosowanie, obejmują dostępność, poziom szczegółowości i jakość danych wejściowych wymaganych przez modele, adekwatność rozpoznanych powiązań oraz zdolność modelu do generowania odpowiedniego zakresu wyników wokół oczekiwanych wartości;
- Model spełnia specyfikacje; oraz
- Pełne lub częściowe wyniki modelu są powtarzalne lub czy jakiekolwiek różnice są objaśnialne.

Walidacja modelu powinna być przeprowadzona przez osobę (osoby), która nie tworzyła modelu, chyba że spowoduje to obciążenie niewspółmierne do ryzyka modelu.

c) Aktuariusz powinien wykorzystując wyniki przebiegu modelu:

- Być przekonanym, że spełnione zostały warunki zastosowania modelu;
- Być przekonanym, że do danych wejściowych i wyjściowych są zastosowane odpowiednie kontrole;
- Rozważyć, czy walidację modelu opisaną w pkt. b) należy przeprowadzić w całości czy częściowo;
- Rozumieć, a w stosownych przypadkach, wyjaśnić istotne różnice między różnymi uruchomieniami modelu i być przekonanym, że istnieje odpowiedni proces kontroli uruchomień modelu. W przypadku modeli stochastycznych aktuariusz stosujący model powinien się upewnić, że wykonano wystarczającą liczbę uruchomień modelu i rozumieć znaczące różnice między poszczególnymi przebiegami modelu;
- Rozumieć wszystkie działania zarządu lub reakcje zarządcze przyjęte w modelu. We wszystkich raportach aktuariusz powinien ujawnić takie zakładane działania zarządu lub reakcje zarządcze (ang. management actions) przyjęte w modelu oraz ich szeroko rozumiane implikacje.
- Udokumentować, w stosownych przypadkach, ograniczenia, dane wejściowe, kluczowe założenia, cel zastosowania i wyniki modelu.

Zadanie 8.

a) (**3*p.***) Przedstaw ideę i sposób konstrukcji wykresów PDP (*Partial Dependence Plot*).

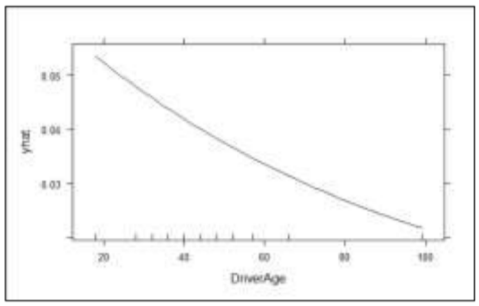
b) (**2*p.***) Liczbę roszczeń K_i w pewnym portfelu ubezpieczeń AC modelowano z uwzględnieniem następujących zmiennych objaśniających:

DriverAge – wiek kierowcy (w latach),
CarAge - wiek samochodu (w latach),
Brand – marka samochodu. Zmienna jakościowa przyjmująca następujące kategorie: *A, B, C, D, E, F* i *G*.

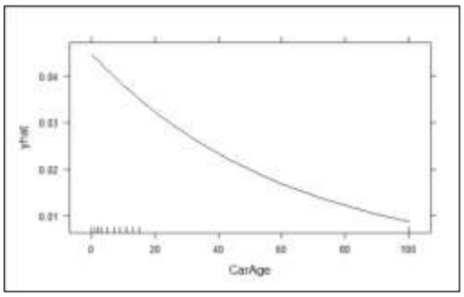
Oszacowano uogólniony model liniowy, w którym przyjęto rozkład Poissona dla K_i oraz link kanoniczny. Dla tego modelu skonstruowano wykresy PDP przedstawione na rysunku 8.1. Podaj interpretację tych wykresów.

Rys. 8.1

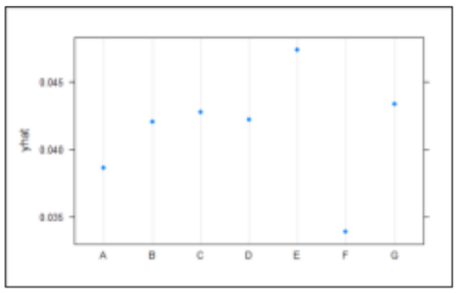
Wiek kierowcy



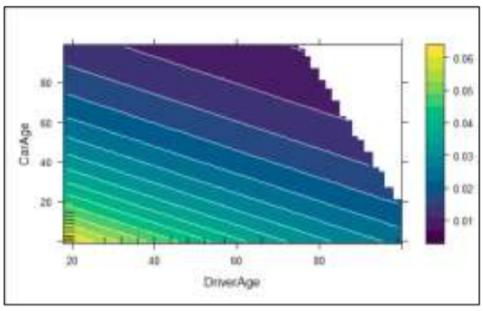
Wiek samochodu



Marka samochodu



Wiek kierowcy i samochodu



a) Główną ideą wykresów PDP jest pokazanie, jak zmiana wartości jednej lub dwóch wybranych cech wpływa na średnią predykcję modelu, przy jednoczesnym uśrednieniu efektów wszystkich pozostałych cech. Innymi słowy, wykres PDP ilustruje **efekt krańcowy** (marginalny) wybranej cechy na prognozę modelu.

Pozwala to na zrozumienie, czy zależność między cechą a predykcją jest liniowa, monotoniczna, czy bardziej złożona, co jest szczególnie przydatne w przypadku modeli takich jak lasy losowe czy sieci neuronowe.

Konstrukcja wykresu częściowej zależności dla pojedynczej cechy x_S przebiega w następujących krokach:

1. **Wybór siatki wartości**: Dla analizowanej cechy x_S wybiera się zbiór interesujących wartości (tzw. siatkę), dla których będzie badany jej wpływ.
2. **Modyfikacja danych**: Dla każdej obserwacji w zbiorze danych (np. uczącym) i dla każdej wartości z siatki:
   - Sztucznie ustawia się wartość cechy x_S na daną wartość z siatki.

- Wartości wszystkich pozostałych cech ($x_{\bar{S}}$) pozostawia się bez zmian.

3. **Predykcja**: Model generuje predykcję dla każdej tak zmodyfikowanej obserwacji.
4. **Uśrednianie**: Oblicza się średnią ze wszystkich predykcji uzyskanych w poprzednim kroku. Wynik jest pojedynczym punktem na wykresie PDP, odpowiadającym jednej wartości z siatki dla cechy x_S.
5. **Wizualizacja**: Powtarza się kroki 2-4 dla wszystkich wartości z siatki, a następnie tworzy wykres, na którym oś X reprezentuje wartości cechy x_S, a oś Y – odpowiadające im średnie predykcje.

b) • Zmienne DriverAge, CarAge mają nieliniowy wpływ na prognozowaną liczbę szkód.

• Zmienna Brand ma wpływ na prognozę w zależności od wartości.

• Między zmiennymi nie ma silnych efektów interakcji co sugeruje prawie równoległy przebieg linii konturowych na wykresie konturowym.

Zadanie 9.

a) (***3p.***) Przedstaw ideę i konstrukcję testu ilorazu wiarygodności. Zapisz hipotezę zerową i alternatywną i wskaż czy różnią się one od hipotez (zerowej i alternatywnej) stawianych w testach zgodności (np. chi-kwadrat, Kołmogorowa-Smirnowa). Podaj postać statystyki testowej i jej rozkład.

b) (***2p.***) Wiadomo, że wysokość szkód w pewnym portfelu ubezpieczeń ma rozkład Pareto z parametrem $\alpha = 2$ i nieznanym parametrem θ. Z portfela wylosowano 20 szkód i oszacowano θ metodą największej wiarygodności, uzyskując wartość 7.0 ($\hat{\theta} = 7.0$). Następnie z wykorzystaniem testu ilorazu wiarygodności testowano hipotezę zerową H_0: $\theta = 3.1$. Wyznacz prawdopodobieństwo testowe (*p-Value*) dla tego testu.

Uwaga! $\sum_{i=1}^{20} \ln(x_i + 7.0) = 49.01$; $\sum_{i=1}^{20} \ln(x_i + 3.1) = 39.30$

Funkcja gęstości rozkładu Pareto ma postać: $f(x) = \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}}$.

a)

Test ilorazu wiarygodności jest narzędziem statystycznym służącym do porównywania dwóch konkurencyjnych modeli (rozkładów prawdopodobieństwa) w celu określenia, który z nich lepiej pasuje do obserwowanych danych.

Podstawowa idea polega na tym, że jeśli model bardziej złożony (hipoteza alternatywna) jest prawdziwy, to wymuszenie dopasowania prostszego modelu (hipoteza zerowa) powinno skutkować znacząco niższą wartością funkcji wiarygodności.

Hipotezy:

H_0: Dane pochodzą z populacji o rozkładzie A (model prostszy).

H_1: Dane pochodzą z populacji o rozkładzie B (model bardziej złożony).

Kluczowym warunkiem jest, aby model w hipotezie zerowej był zagnieżdżony w modelu z hipotezy alternatywnej. Oznacza to, że model A można otrzymać z modelu B przez nałożenie ograniczeń na jego parametry (np. model wykładniczy jest szczególnym przypadkiem modelu gamma, gdy jeden z parametrów jest ustalony na 1).

Hipotezy w teście ilorazu wiarygodności różnią się od tych w testach zgodności, takich jak test chi-kwadrat czy test Kołmogorowa-Smirnowa. Testy zgodności sprawdzają, czy dane pochodzą z jednego, konkretnego rozkładu:

H_0: Dane pochodzą z populacji o danym rozkładzie.

H_1: Dane nie pochodzą z populacji o danym rozkładzie.

Statystyka testowa:

$$T = 2\ln\left(\frac{L_1}{L_0}\right) = 2(\ln L_1 - \ln L_0),$$

gdzie L_1, L_2 to wiarygodności modeli z hipotezy zerowej i alternatywnej.

Dla dużych prób statystyka testowa T ma w przybliżeniu rozkład chi-kwadrat (χ^2). Liczba stopni swobody tego rozkładu jest równa różnicy w liczbie wolnych (niezależnych) parametrów między modelem z hipotezy alternatywnej a modelem z hipotezy zerowej.

b) Funkcja wiarygodności:

$$L(\alpha, \theta; x_i) = \prod_{i=1}^{20} \frac{\alpha \theta^{\alpha}}{(x_i + \theta)^{\alpha+1}} = \frac{\alpha^{20} \theta^{20\alpha}}{\prod_{i=1}^{20} (x_i + \theta)^{\alpha+1}}$$

$$\ln[L(\alpha, \theta; x_i)] = 20 \ln(\alpha) + 20\alpha \ln(\theta) - (\alpha+1) \sum_{i=1}^{20} \ln(x_i + \theta)$$

Dla H_0:

$$\ln(L_0) = 20 \ln(2) + 40 \ln(3.1) - 3 \sum_{i=1}^{20} \ln(x_i + 3.1) = -58.7810$$

Dla H_1:

$$\ln(L_1) = 20 \ln(2) + 40 \ln(7) - 3 \sum_{i=1}^{20} \ln(x_i + 7) = -55.3307$$

$$T = 2(\ln(L_1) - \ln(L_2)) = 2 \cdot [-55.3307 - (-58.3307)] = 6.901$$

T ma rozkład χ^2_1

p-value = 0.0086 co można odczytać z tabeli wartości $\chi^2_{\alpha; \nu}$

Zadanie 10.

a) (**2*p.***) Wskaż co najmniej cztery reguły określające, kiedy węzeł w drzewie regresyjnym jest przyjmowany za końcowy (jest uznawany za liść).

b) (**1*p.***) Na czym polega i w jakim celu stosuje się przycinanie drzewa regresyjnego?

c) (**2*p.***) Liczbę roszczeń K_i w pewnym portfelu ubezpieczeń AC modelowano z uwzględnieniem następujących zmiennych objaśniających:

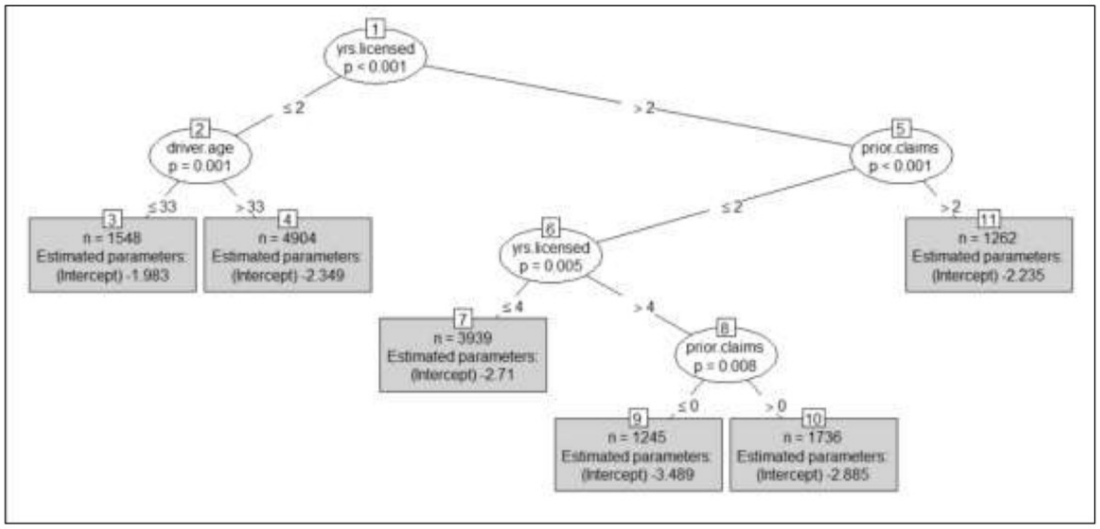
driver.age – wiek kierowcy w latach (zmienna ilościowa),

prior.claims – liczba wcześniej zgłoszonych roszczeń (zmienna ilościowa),

yrs.licensed – okres posiadania prawa jazdy w latach (zmienna ilościowa).

Przyjęto dla K_i rozkład Poissona i skonstruowano binarne drzewo GLM (*Generalized Linear Model Tree*) przedstawione na rysunku 10.1. Dla liści podano oszacowania modeli regresji Poissona z linkiem kanonicznym. Opisz grupę kierowców, która średnio rocznie zgłasza najwięcej szkód i grupę, która średnio rocznie zgłasza najmniej szkód. Oszacuj dla tych grup prawdopodobieństwa wystąpienia co najmniej jednego roszczenia.

Rys. 10.1



a) Węzeł jest uznawany za końcowy:

- jeśli zawiera mniej niż z góry określoną liczbę obserwacji.
- gdy jego głębokość (czyli odległość od korzenia drzewa) osiągnie ustalony limit.
- jeśli w wyniku tego podziału co najmniej jeden z nowo powstałych węzłów potomnych (liści) zawierałby mniej niż z góry określoną liczbę obserwacji.
- jeśli najlepszy możliwy podział tego węzła nie przynosi spadku dewiancji (miary błędu) o wartość większą niż ustalony próg.

b) Przycinanie polega budowaniu maksymalnie rozbudowanego drzewa, pozwalając mu rosnąć aż do momentu, gdy dalsze podziały nie są możliwe (np. w liściach zostaje zbyt mało obserwacji lub wszystkie mają tę samą wartość). Takie drzewo jest bardzo złożone i idealnie dopasowane do danych treningowych. Następnie, w sposób systematyczny, usuwa się (przycina) całe gałęzie drzewa, czyli węzły wraz z ich potomkami. Celem jest znalezienie optymalnego poddrzewa, które stanowi najlepszy kompromis między prostotą a dokładnością predykcji.

c) Średnia roczna liczba zgłaszanych szkód jest najwyższa dla grupy o najwyższej wartości parametru Intercept i najniższa dla grupy o najniższej wartości tego parametru. Wynika to z linku kanonicznego:

$$\mu = \exp\{\beta_0\}$$

Dla liścia nr 3:

$$\mu = \exp(-1.983) = 0.13765565$$

$$P(K=k) = \frac{1}{k!} e^{-\lambda} \lambda^k \quad (\mu = \lambda \text{ w modelu Poissona})$$

$$P(K \geq 1) = 1 - P(K=0) = 1 - e^{-\mu} = 0.12860194$$

Najwięcej szkód powodują kierowcy, którzy posiadają prawo jazdy nie dłużej niż 2 lata i mają nie więcej niż 33 lata (niezależnie od liczby zgłoszonych wcześniej szkód)

Dla liścia nr 9:

$$\mu = \exp\{-3.489\} = 0.030531388$$

$$P(K \geq 1) = 0.030070013$$

Najmniej szkód powodują kierowcy, którzy posiadają prawo jazdy powyżej czterech lat i wcześniej nie zgłosili żadnej szkody (niezależnie od wieku).