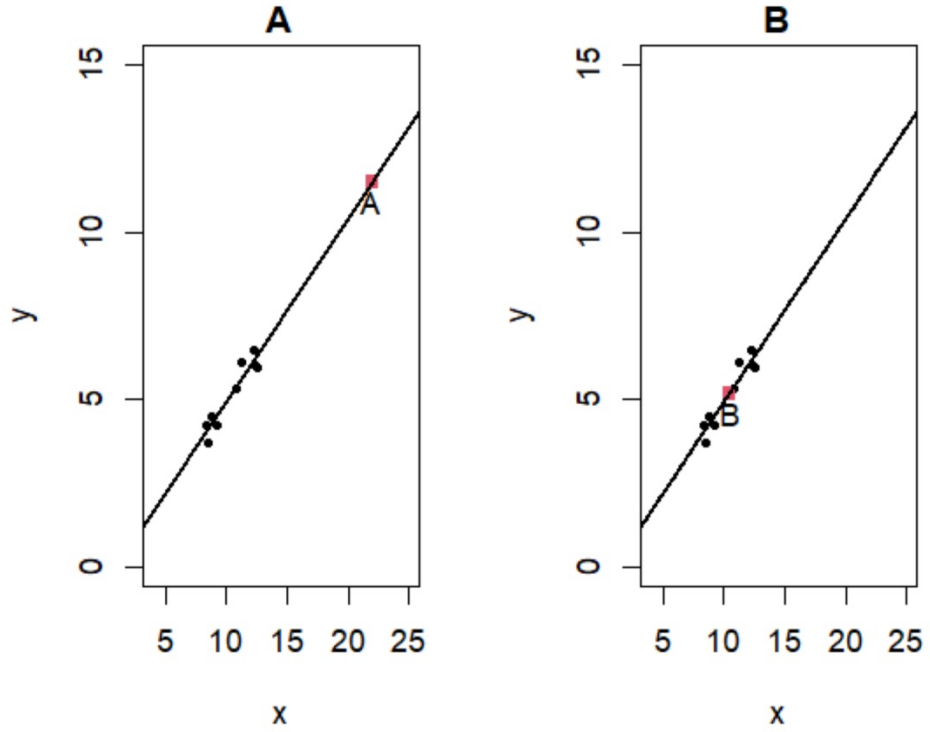
Zadanie 1.

Na wykresach zamieszczonych na poniższym rysunku (Rys.1.1) linie ciągłe przedstawiają proste regresji wyznaczone w oparciu o ten sam zbiór danych *Z*, do którego nie należą wyróżnione obserwacje A i B (kwadraty w kolorze czerwonym), tj. na wykresie A (lewym) bez obserwacji A, a na wykresie B (prawym) bez obserwacji B. Zarówno punkt A, jak i punkt B leżą na prostej regresji.

Rys.1.1.



a) ***(1p.)*** Podaj definicję obserwacji wpływowej (*influential observation*).

b) ***(2p.)*** Opisz jedną z metod identyfikacji obserwacji wpływowych.

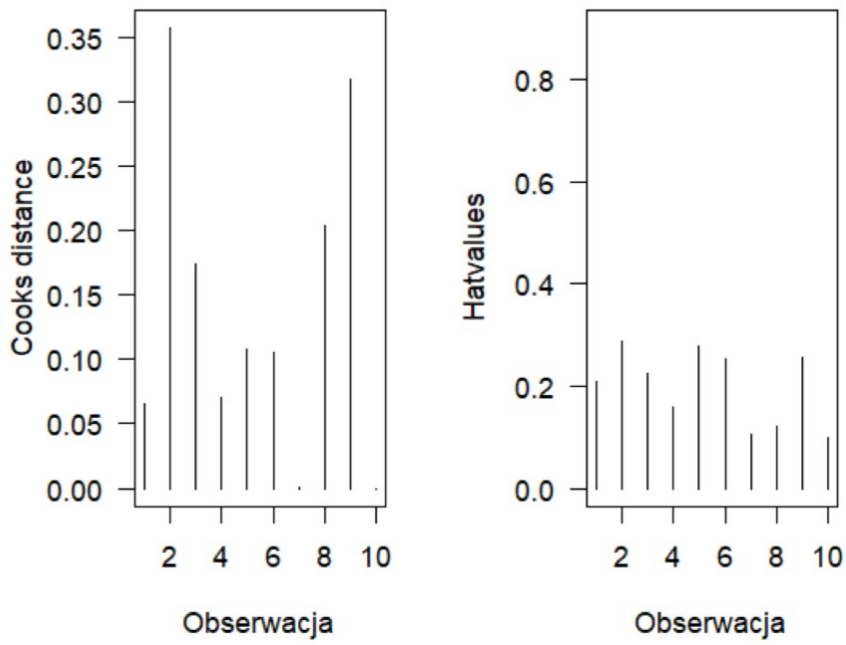
c) ***(2p.)*** Oszacowano dwa modele regresji liniowej:

   a. model A – w oparciu o zbiór danych przedstawiony na rys. 1.1A, tj. zbiór *Z* łącznie z obserwacją A,

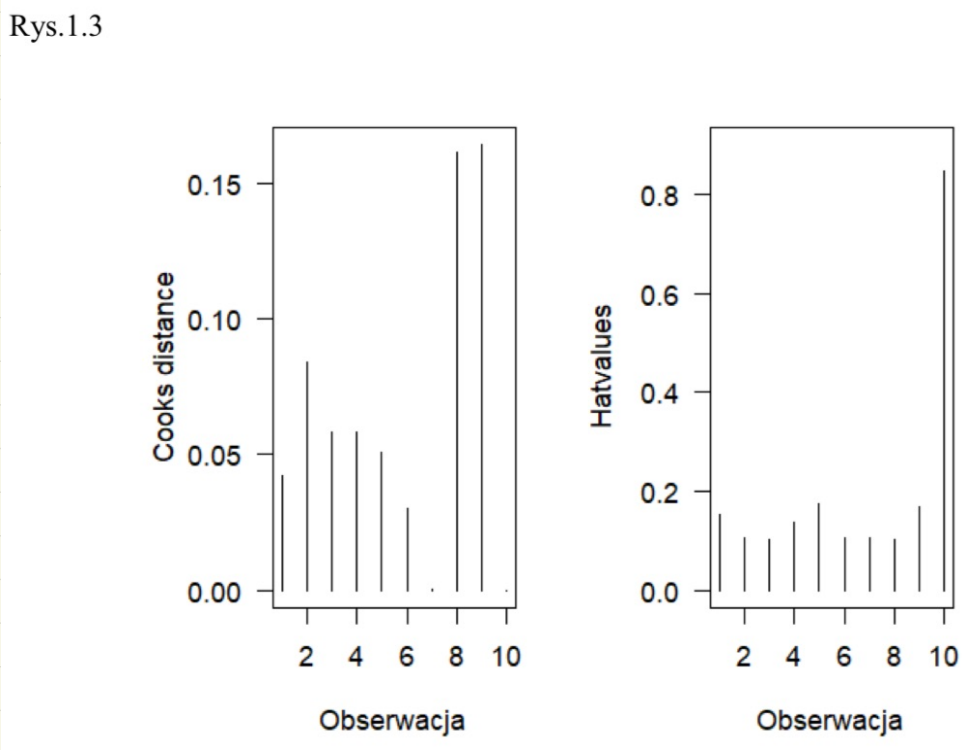
   b. model B – w oparciu o zbiór danych przedstawiony na rys. 1.1B, tj. zbiór *Z* łącznie z obserwacją B,

   Na rysunkach 1.2 i 1.3 przedstawiono wykresy odległości Cooka (*Cook's Distance*) i dźwigni (*leverages*) dla tych modeli. Wskaż i **uzasadnij**, który z rysunków przedstawia wykresy odpowiadające modelowi A, a który modelowi B. **Uwaga! Na wykresach obserwacja A i obserwacja B ma numer 10.**

Rys.1.2



Rys.1.3



a) Obserwacja wpływowa to taka obserwacja, która, jeśli zostanie nieznacznie zmieniona lub pominięta, w istotny sposób zmodyfikuje oszacowania parametrów modelu.

b) Odległość Cooka mierzy odległość między $\hat{\beta}$ a $\hat{\beta}_{(-i)}$, (ostatnie to wektor dopasowanych współczynników na zbiore bez i-tej obserwacji) przy czym duże wartości wskazują na obserwacje, które mają wpływ na oszacowane współczynniki regresji. Odległość Cooka służy do zbadania, jak każda obserwacja wpływa na cały wektor parametrów $\hat{\beta}$.

c) Model A - rys. 1.3

Model B - rys. 1.2

Dźwignia (Hatvalue) jest miarą tego, jak bardzo wartość zmiennej objaśniającej dla danej obserwacji odbiega od średniej wartości tej zmiennej dla całego zbioru danych. Obserwacja 4 odbiega od reszty obserwacji na rys. 1.1, wysoka wartość hatvalue dla tej obserwacji jest na rys. 1.3.

Zadanie 2.

Opracowano dwa plany taryfikacji: **Plan A** i **Plan B.** Plany te przetestowano na zbiorze liczącym 5 obserwacji. Uzyskano następujące wyniki (tab. 2.1):

Tab. 2.1

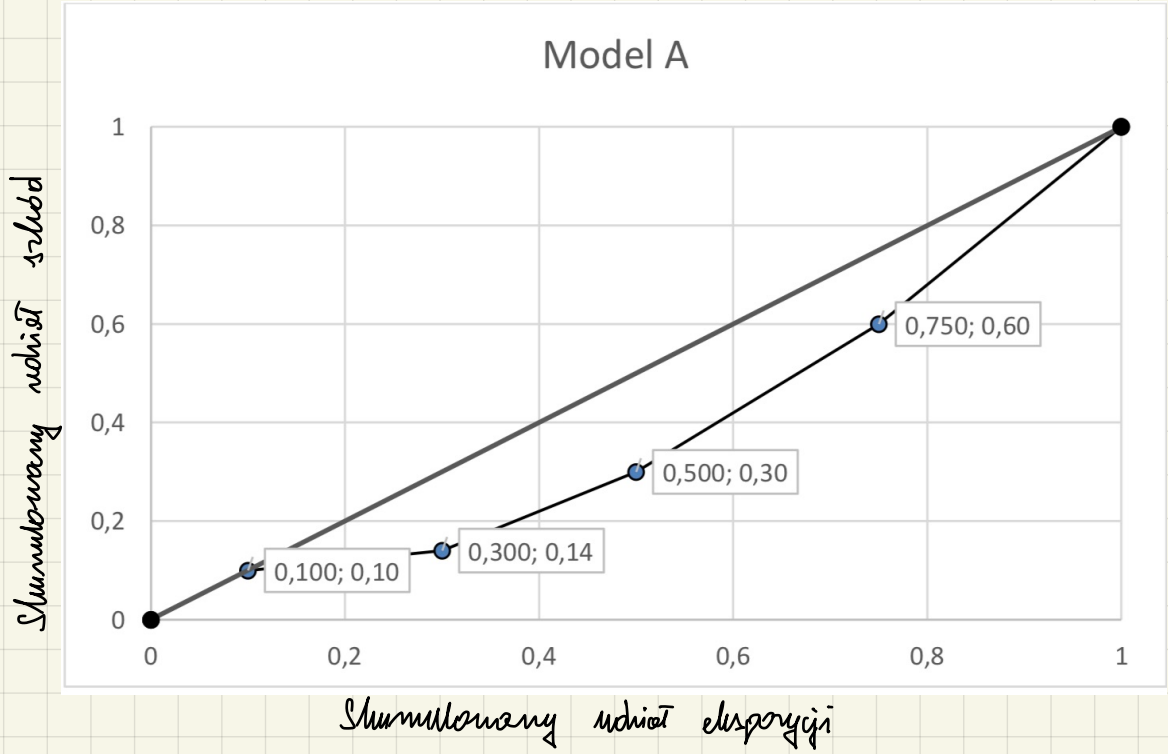
Ryzyko	Ekspozycja	Prognozy		Wartości rzeczywiste
		Plan A	Plan B	
1	1	21	22	20
2	0.4	3	6	5
3	1	15	15	15
4	0.8	9	5	8
5	0.8	4	3	2

a) ***(2p.)*** Narysuj krzywą Lorenza dla **planu A**.

b) ***(3p.)*** Obliczyć współczynnik Giniego dla planu A i B, wiedząc, że pole pod krzywą Lorenza dla **planu A** wynosi: 0.3855, a dla **planu B**: 0.3655. W oparciu o uzyskane wyniki wskaż lepszy plan. Wybór uzasadnij!

a) W pierwszym kroku należy uporządkować dane rosnąco według prognoz

Ryzyko	Ekspozycja	Prognoza	Rzeczywista	Kumulanty Ekspozycji	Kumulanty Rzeczywistych	Udział Kumulant Ekspozycji	Udział Kumulant Rzeczywistych
2	0.4	3	5	0.4	5	$\frac{0.4}{4} = 0.1$	$\frac{5}{50} = 0.1$
5	0.8	4	2	1.2	7	$\frac{1.2}{4} = 0.3$	$\frac{7}{50} = 0.14$
4	0.8	9	8	2	15	$\frac{2}{4} = 0.5$	$\frac{15}{50} = 0.3$
3	1	15	15	3	30	$\frac{3}{4} = 0.75$	$\frac{30}{50} = 0.6$
1	1	21	20	4	50	$\frac{4}{4} = 1$	$\frac{50}{50} = 1.0$



b) Współczynnik Giniego:

$G = 1 - 2 \cdot$ (pole pod krzywą Lorenza)

$G_A = 1 - 2 \cdot 0.3855 = 0.229$

$G_B = 1 - 2 \cdot 0.3655 = 0.269$

Im wyższa wartość współczynnika Giniego, tym model lepiej różnicuje ryzyko, czyli skuteczniej oddziela klientów o niskiej szkodowości od tych o wysokiej.

Preferowany jest plan B.

Zadanie 3.

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 3.1 (w nawiasach podano ekspozycję na ryzyko):

Tab. 3.1

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

a) ***(2p.)*** Twoim zadaniem jest oszacowanie, w oparciu o te dane, modelu regresji Poissona (z linkiem kanonicznym), z uwzględnieniem obydwu zmiennych objaśniających. Zakoduj zmienne objaśniające, przyjmując jako kategorie referencyjne: K i D1. Podaj:

- postać macierzy modelu (*model matrix*),
- wektor zmiennej zależnej (obserwacji)
- wektor przedstawiający ekspozycję na ryzyko.

b) ***(3p.)*** Po oszacowaniu modelu dysponujesz następującymi wynikami:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.7174     0.1898 -14.318  < 2e-16 ***
PlecM        -0.2853     0.1631  -1.750  0.08014 .
DystansD2     0.5141     0.1887   2.725  0.00644 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 22.75602  on 3  degrees of freedom
Residual deviance:  0.45335  on 1  degrees of freedom
AIC: 28.706
```

Wyznacz resztę Pearsona (*Pearson residual*) dla kobiet (*K*), które w ciągu roku przebyła dystans mniejszy niż 20 tys. km (*D1*).

Wyjaśnij, dlaczego w przypadku diagnostyki uogólnionych modeli liniowych korzystniejsze jest stosowanie reszt Pearsona zamiast zwykłych reszt (*raw residuals*, *response residuals*).

a) • Macierz modelu:

Należy zakodować zmienne płci i dystansu zmiennymi zero-jedynkowymi oraz uwzględnić wyraz wolny.

(intercept)	PlecM	DystansD2
1	0	0
1	0	1
1	1	0
1	1	1

Plec = K $\Rightarrow x_1 = 0$

Plec = M $\Rightarrow x_1 = 1$

Dystans = D1 $\Rightarrow x_2 = 0$

Dystans = D2 $\Rightarrow x_2 = 1$

- Wektor zm. zależnej:

L. szkod
15
197
28
35

- Wektor przedstawiający ekspozycję na ryzyko

Ekspozycja
200
1800
600
400

b) Reszta Pearsona:

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

Dla rozkładu Poissona, gdzie wariancja jest równa wartości oczekiwanej:

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

$\hat{\mu}_i$ - można obliczyć z linku kanonicznego uwzględniając ekspozycję:

$$\hat{\mu}_1 = E_1 \cdot \exp\{\beta_0 + \beta_1 X_1 + \beta_2 X_2\}$$

W modelu: Płeć = K => $X_1 = 0$

Dystans = D1 => $X_2 = 0$

$$\hat{\mu}_1 = E_1 \cdot \exp\{\beta_0\} = 200 \cdot \exp\{-2.7174\} = 13.20936$$

$$r_1^P = \frac{15 - 13.20936}{\sqrt{13.20936}} = 0.4926841$$

Reszty Pearsona są preferowane, ponieważ poprzez normalizację (dzielenie zwykłych reszt przez odchylenie standardowe i wagę) "stabilizują" wariancję reszt. Dzięki temu są one bardziej porównywalne w całym zakresie wartości dopasowanych i pozwalają na trafniejszą ocenę dopasowania modelu oraz identyfikację obserwacji odstających, bez zakłóceń wynikających z naturalnej zależności wariancji od średniej w rodzinie rozkładów wykładniczych.

Zadanie 4.
Na podstawie 40-stu obserwacji zebrano następujące dane (symbol (*) oznacza obserwacje cenzurowane z góry):
4 4 4 4* 5* 5* 5* 5* 5* 6 6 6 6 6 6* 7* 8* 9 9 9 9 9 9 9* 10* 10* 12* 13 13 13 13 13* 14* 14* 15 15 16* 16* 17* 18*
Niech $S(\cdot)$ oznacza funkcję przeżycia. Wykorzystując te dane:

a) (**2*p.***) Wyznacz estymator Kaplana–Meiera dla $S_{40}(6)$.
b) (**2*p.***) Oszacuj wariancję estymatora Kaplana–Meiera dla $S_{40}(6)$.
c) (**1*p.***) Skonstruuj 95% przedział ufności dla $S(6)$. Wykorzystaj estymację Kaplana–Meiera.

a) Estymator Kaplana - Meiera :

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

i	y_i	d_i	b_i	r_i	$S_n(y_i)$
1	4	3	1	40	$1 - \frac{3}{40} = \frac{37}{40}$
2	6	5	3	31	$\frac{37}{40} \cdot \left(1 - \frac{5}{31}\right) = \frac{37}{40} \cdot \frac{26}{31} = 0.77581$

$S_{40}(6) = 0.77581$

b) Do obliczenia wariancji używamy wzoru Greenwooda :

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

$$Var(S_{40}(6)) = 0.77581^2 \cdot \left[\frac{3}{40 \cdot 37} + \frac{5}{31 \cdot 26}\right] = 0.004953783$$

c) Przedział ufności :

$$S_n(y) \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{Var(S_n(y))}$$

$z_{0.975} = 1.96$

$$0.77581 \pm 1.96 \sqrt{0.004953783} = [0.637859, 0.913761]$$

Zadanie 5.

Szereg czasowy liczący 123 obserwacje podzielono na zbiór uczący liczący 120 obserwacji ($t = 1, \dots, 120$) i testowy z trzema obserwacjami ($t = 121, 122, 123$). Na podstawie zbioru uczącego rozważano kilka różnych specyfikacji modelu klasy ARIMA i ostatecznie wybrano i oszacowano następujący model AR(1):

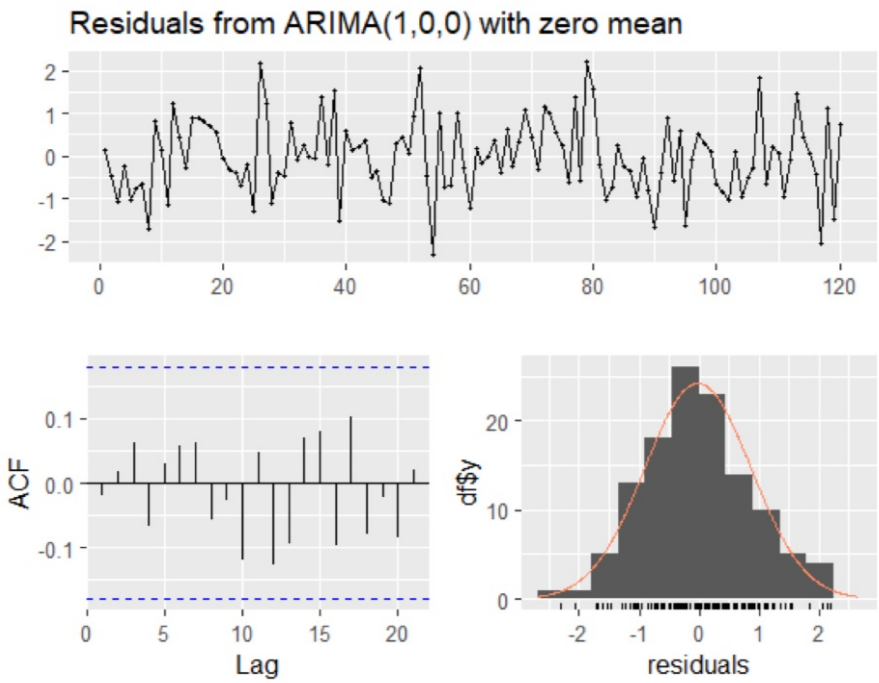
```
ARIMA(1,0,0) with zero mean
Coefficients:
         ar1
      0.6913
s.e.  0.0647

sigma^2 = 0.7881:  log likelihood = -155.8
AIC=315.61   AICc=315.71   BIC=321.18
```

Wartości szeregu dla 4 ostatnich okresów są równe:

t	*120*	*121*	*122*	*123*
y_t	-0.40	-0.30	-0.18	-0.12

a) (**2*p.***) Dysponując następującymi wykresami związanymi z analizą reszt oszacowanego modelu (rys. 5.1), oceń poprawność wybranej specyfikacji.

Rys.5.1



b) (**1*p.***) Wyznacz prognozy szeregu na okresy ze zbioru testowego.

c) (**2*p.***) Oceń dokładność wyznaczonych prognoz wykorzystując miernik *MAPE* (*Mean Absolute Percentage Error*).

a) • Szereg czasowy reszt – wykres reszt modelu w funkcji czasu nie wykazuje żadnych widocznych wzorców. Reszty oscylują losowo wokół zera, co jest pożądaną cechą i wskazuje na stacjonarność reszt.

• Funkcja autokorelacji reszt (ACF) – wszystkie współczynniki autokorelacji dla różnych opóźnień (Lag) mieszczą się wewnątrz przedziału ufności (niebieskie linie). Oznacza to, że są one statystycznie nieistotne.

Brak istotnych autokorelacji w resztach świadczy o tym, że model dobrze wychwycił zależność sezonową obecną w oryginalnych danych, a reszty mają charakter białego szumu.

- Histogram reszt – ma kształt zbliżony do dzwonowatego, a nałożona na niego krzywa gęstości rozkładu normalnego dobrze go przybliża. Sugeruje to, że założenie o normalności błędów losowych jest spełnione.

b) $y^p_{121} = 0.6913 \cdot (-0.40) = -0.27652$

$y^p_{122} = 0.6913 \cdot (-0.27652) = -0.1911583$

$y^p_{123} = 0.6913 \cdot (-0.1911583) = -0.1321477$

c)

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

$$MAPE = \frac{1}{3} \left[\left| \frac{-0.30 + 0.27652}{-0.30} \right| + \left| \frac{-0.18 + 0.1911583}{-0.18} \right| + \left| \frac{-0.12 + 0.1321477}{-0.12} \right| \right] =$$

$$= 0.07840323 2$$

Zadanie 6.

a) (***3p.***) Zmienna losowa X ma standardowy rozkład normalny ($X \sim N(0,1)$). Wykorzystując przesunięty rozkład wykładniczy z parametrem $\lambda = 1$ i odpowiednio dobranym parametrem przesunięcia, wyznacz estymator Monte Carlo wykorzystujący próbkowanie ważone (*importance sampling estimator*) dla prawdopodobieństwa $P(X > 3.5)$.

b) (***2p.***) Za pomocą tego estymatora oszacuj $P(X > 3.5)$, wykorzystując następującą próbkę wylosowaną z rozkładu jednostajnego (na przedziale $(0,1)$): 0.8880, 0.4704, 0.11076.

Uwaga! Funkcja gęstości i dystrybuanta rozkładu wykładniczego są następujące:

$$f(x) = \lambda e^{-\lambda x}, \;\; F(x) = 1 - e^{-\lambda x}, x \geq 0$$

a) Estymator:

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

$f(x) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}x^2\}$ - rozkład $N(0,1)$

$g(x) = \exp\{-(x-3.5)\}$ - przesunięty rozkład wykładniczy z $\lambda = 1$

$h(x) = \mathbb{1}_{x>3.5} = 1$ - funkcja zawsze równa 1 bo losujemy z rozkładu przesuniętego o 3.5

Po wstawieniu do wzoru:

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

b) Trzeba wygenerować próbkę z przesuniętego rozkładu wykładniczego. Mamy podaną próbkę z rozkładu jednostajnego więc możemy skorzystać z metody odwrotnej dystrybuanty.

$y = 1 - \exp\{-x+3.5\}$

$y - 1 = -\exp\{-x+3.5\}$

$1 - y = \exp\{-x+3.5\}$

$3.5 - x = \ln(1-y)$

$$-x = -3.5 + m(1-y)$$

$$x = 3.5 - m(1-y)$$

$$x_1 = 3.5 - m(1 - 0.8880) = 5.689256$$

$$x_2 = 3.5 - m(1 - 0.4704) = 4.135633$$

$$x_2 = 3.5 - m(1 - 0.11076) = 3.617388$$

$$\hat{P}(X > 3.5) = \frac{1}{3\sqrt{2\pi}} \Big[\exp\{-\tfrac{1}{2} \cdot 5.689256^2 + 5.689256 - 3.5\} +$$

$$+ \exp\{-\tfrac{1}{2} \cdot 4.135633^2 + 4.135633 - 3.5\} +$$

$$+ \exp\{-\tfrac{1}{2} \cdot 3.617388^2 + 3.617388 - 3.5\}\Big] =$$

$$= 0.0002640478$$

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

a) (**2*p.***) walidacji modeli,
b) (**3*p.***) walidacji i braku danych.

Walidacja modelu

Walidacja modelu obejmuje ocenę, czy:

- **Model jest dopasowany do zamierzonego celu pracy**. W tym kontekście aktuariusz powinien wziąć pod uwagę takie aspekty jak:
  - Dostępność, jakość i poziom szczegółowości danych wejściowych wymaganych przez model.
  - Adekwatność powiązań rozpoznanych w modelu.
  - Zdolność modelu do generowania odpowiedniego zakresu wyników wokół oczekiwanych wartości.
- **Model spełnia swoje specyfikacje**.
- **Wyniki modelu (pełne lub częściowe) są powtarzalne**, a wszelkie ewentualne różnice dają się wyjaśnić.

Standard wskazuje, że walidacja modelu **powinna być przeprowadzona przez osobę lub osoby, które nie tworzyły danego modelu**. Odstępstwo od tej zasady jest możliwe tylko wtedy, gdyby jej zastosowanie powodowało obciążenie niewspółmierne do ryzyka związanego z modelem. Dodatkowo, przy wykorzystywaniu wyników z konkretnego uruchomienia modelu, aktuariusz powinien rozważyć, czy walidacja nie powinna zostać przeprowadzona ponownie w całości lub w części.

Walidacja Danych

Aktuariusz powinien podjąć uzasadnione kroki w celu sprawdzenia spójności, kompletności i dokładności wykorzystywanych danych. Możliwe działania obejmują między innymi:

- **Uzgodnienie z dokumentami finansowymi**: Porównanie danych z zaudytowanymi sprawozdaniami finansowymi, zestawieniami obrotów i sald lub innymi odpowiednimi dokumentami, jeśli są dostępne.
- **Testowanie racjonalności**: Porównanie danych z danymi zewnętrznymi lub niezależnymi w celu oceny ich racjonalności.
- **Sprawdzenie spójności wewnętrznej**: Przetestowanie danych pod kątem ich wewnętrznej spójności oraz spójności z innymi istotnymi informacjami.
- **Porównanie z danymi historycznymi**: Zestawienie danych z danymi za poprzedni okres lub okresy.

Aktuariusz jest zobowiązany opisać podjęte kroki walidacyjne we wszystkich tworzonych raportach.

Postępowanie w Przypadku Braku Danych

W przypadku stwierdzenia braków w danych (takich jak nieadekwatność, niespójność czy niekompletność), aktuariusz musi wziąć pod uwagę ich możliwy wpływ na wyniki pracy.

1. **Ocena istotności**: Jeżeli braki w danych prawdopodobnie nie będą miały istotnego wpływu na wyniki, nie muszą być dalej rozpatrywane.
2. **Postępowanie przy istotnych brakach**: Jeżeli aktuariusz nie jest w stanie w zadowalający sposób usunąć istotnych braków, powinien rozważyć jedną z poniższych opcji:
    - **Odmówić lub przerwać świadczenie usług** zawodowych.
    - **Współpracować ze zleceniodawcą** w celu modyfikacji zlecenia lub uzyskania dodatkowych, odpowiednich danych.
    - **Wykonać usługi w najlepszy możliwy sposób**, jednocześnie ujawniając we wszystkich raportach informacje o brakach w danych oraz wskazując ich potencjalny wpływ na wyniki.

Zadanie 8.

Zanotowano następujące kwoty roszczeń: 130, 30, 90, 60, 190. Dopasowano do nich rozkład wykładniczy, dla którego średnią oszacowano metodą momentów.

a) (***2p.***) Skonstruuj wykres prawdopodobieństwo-prawdopodobieństwo (*p-p plot, probability plot*) i na jego podstawie oceń czy wybrano właściwy model.

b) (***3p.***) Sprawdź dopasowanie tego rozkładu testem Kołmogorowa-Smirnowa. Przyjmij poziom istotności 0.05. Wartość krytyczna dla tego poziomu istotności wynosi: $\frac{1.36}{\sqrt{n}}$, gdzie n oznacza liczbę obserwacji.

Uwaga ! Dla rozkładu wykładniczego: $F(x) = 1 - e^{-\lambda x}, x \geq 0, E(X) = \frac{1}{\lambda}$

Średnią oszacowano metodą momentów dlatego trzeba porównać średnią z próby z wartością oczekiwaną.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{\lambda}$$

$$\hat{\lambda} = \frac{1}{\bar{x}}$$

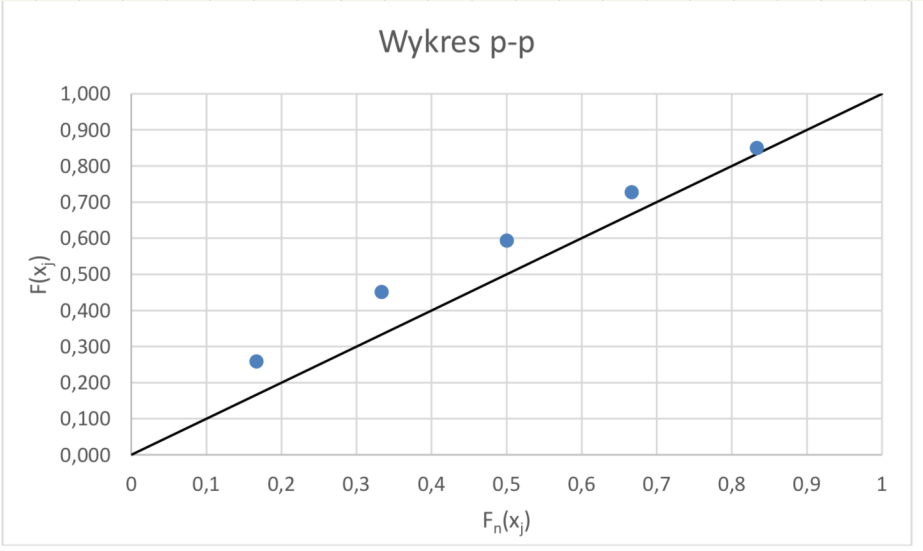
$$\bar{x} = \frac{130+30+90+60+190}{5} = \frac{500}{5} = 100$$

$$\hat{\lambda} = \frac{1}{100} = 0.01$$

$$F(x) = 1 - e^{-0.01x}$$

a) Trzeba obliczyć dystrybuantę empiryczną (w przypadku wykresu p-p bierzemy liczbę obserwacji przez n+1) dla danych uporządkowanych rosnąco.

j	x_j	$\hat{F}_n(x_j)$	$F(x_j)$
1	30	$\frac{1}{6} = 0.167$	$1 - \exp\{-0.01 \cdot 30\} = 0.259$
2	60	$\frac{2}{6} = 0.333$	$1 - \exp\{-0.01 \cdot 60\} = 0.451$
3	90	$\frac{3}{6} = 0.500$	$1 - \exp\{-0.01 \cdot 90\} = 0.593$
4	130	$\frac{4}{6} = 0.667$	$1 - \exp\{-0.01 \cdot 130\} = 0.727$
5	190	$\frac{5}{6} = 0.833$	$1 - \exp\{-0.01 \cdot 190\} = 0.850$



Możemy wstępnie przyjąć, że szkody podlegają rozkładowi wykładniczemu.

b) Test Kołmogorowa-Smirnowa - służy do weryfikacji hipotezy o zgodności rozkładu w próbie z określonym rozkładem teoretycznym.

H_0: Dane pochodzą z rozkładu wykładniczego o parametrze $\lambda = 0.01$

H_1: Dane nie pochodzą z tego rozkładu

Statystyka testowa:

$$D = \max_x |F_n(x) - F(x)|$$

Ponieważ dystrybuanta empiryczna jest funkcją skokową należy porównać dystrybuanty dwa razy, przed i po skoku. Wartości przed skokiem są oznaczone jako $F_n(x-)$. Tym razem dzielimy liczbę obserwacji przez n.

j	x_j	$F_n(x_j -)$	$F_n(x_j)$	$F(x_j)$	$\lvert F_n(x_j -) - F(x_j)\rvert$	$\lvert F_n(x_j) - F(x_j)\rvert$
1	30	0	0.2	0.259	0.259	0.059
2	60	0.2	0.4	0.451	0.251	0.051
3	90	0.4	0.6	0.593	0.193	0.007
4	130	0.6	0.8	0.727	0.127	0.073
5	190	0.8	1	0.850	0.050	0.150

Wartość statystyki to maksimum z dwóch ostatnich kolumn:

$D = 0.259$

Wartość krytyczna:

$$\frac{1.36}{\sqrt{5}} = 0.608$$

$$D = 0.259 < 0.608$$

czyli nie ma podstaw do odrzucenia H_0. Szkody mają rozkład wykładniczy z parametrem $\lambda = 0.01$.

Zadanie 9.

Zadaniem aktuariusza jest wyznaczenie prognoz przedłużenia umowy ubezpieczenia na kolejny rok w pewnym portfelu ubezpieczeń. W tym celu, na podstawie zbioru uczącego, oszacował dwa modele logistyczne: model M1 i model M2. Zmienna zależna w tych modelach przyjmuje dwie wartości: Y=1, gdy klient przedłużył ubezpieczenie oraz Y=0, gdy nie przedłużył). Zmienne objaśniające to:

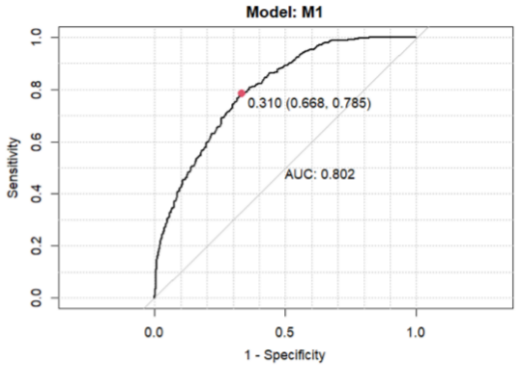
- *plec*: Płeć klienta (K - kobieta, M – mężczyzna)
- *zamieszkanie*: Zamieszkanie klienta (Miasto, Wies)
- *wiek*: Wiek klienta (zmienna jakościowa przyjmująca trzy kategorie: G1, G2, G3, przy czym im wyższa kategoria, tym klienci są starsi).
- *dochod*: Roczny dochód klienta w tys. zł (zmienna ilościowa).

Jakość oszacowanych modeli sprawdził, wykorzystując zbiór testowy.

Na etapie estymacji i testowania modeli uzyskał między innymi następujące wyniki (na wykresach liczby 0.310 i 0.294 oznaczają progi klasyfikacji, dla których uzyskano współrzędne wyróżnionych punktów):

Model M1

Krzywa ROC

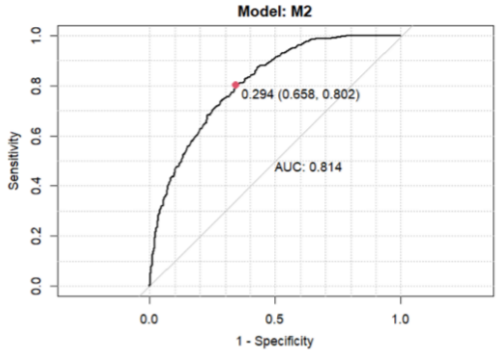


Confusion Matrix
(**dla progu 0.50**)

```
          Reference
Prediction   0   1
         0 722 198
         1 122 202
```

Model M2

Krzywa ROC



Confusion Matrix
(**dla progu 0.50**)

```
          Reference
Prediction   0   1
         0 731 192
         1 113 208
```

Analiza dewiancji

```
Analysis of Deviance Table

Model M1: y ~ plec + zamieszkanie + wiek + dochod
Model M2: y ~ plec + zamieszkanie + wiek + dochod + wiek:dochod
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1859     1894.8
2      1857     1871.6  2   23.141 9.439e-06 ***
```

a) **(2*p*.)** Krótko omów wykorzystanie krzywej ROC do oceny modeli logistycznych.
b) **(2*p*.)** Na podstawie podanych wyników wybierz lepszy model. Wybór uzasadnij.
c) **(1*p*.)** Czy dla wybranego (w podpunkcie b)) modelu zastosowany próg klasyfikacji 0.50 jest optymalny? Odpowiedź uzasadnij.

a) Idealna krzywa ROC "przytula" lewy górny róg wykresu, co oznacza wysoki odsetek prawdziwie pozytywnych wyników przy jednoczesnym niskim odsetku fałszywie pozytywnych wyników. Ogólna wydajność klasyfikatora, zsumowana dla wszystkich możliwych progów, jest podana przez pole pod krzywą ROC (AUC). Idealny klasyfikator ma AUC bliskie 1, podczas gdy klasyfikator, który działa nie lepiej niż losowe zgadywanie, ma AUC równe 0,5.

b) Lepszy jest model M2.

- Analiza deviancji – p-value dla modelu M2 jest bardzo niskie, oznacza to, że dodanie interakcji w modelu M2 przynosi statystycznie istotną poprawę w dopasowaniu modelu w porównaniu z modelem M1.
- Krzywa ROC – wyższa wartość AUC (pole pod wykresem) wskazuje na jego lepszą zdolność do rozróżniania klientów.
- Confusion Matrix:

M1: $(722+202)/1244 = 0.742$

M2: $(731+202)/1244 = 0.754$

Model M2 lepiej klasyfikuje przypadki.

c) Próg 0.50 dla modelu M2:

$$\text{czułość} = \frac{202}{192+202} = 0.520$$

$$\text{specyficzność} = \frac{731}{731+113} = 0.866$$

Próg 0.294 na krzywej ROC dla modelu M2:

czułość = 0.802

specyficzność = 0.706

Próg 0.50 nie jest optymalny, próg na krzywej ROC ma wyższą czułość a w tym przypadku firma woli klasyfikować lepiej klientów, którzy faktycznie przedłużą umowę więc preferuje wyższą czułość.

Zadanie 10.

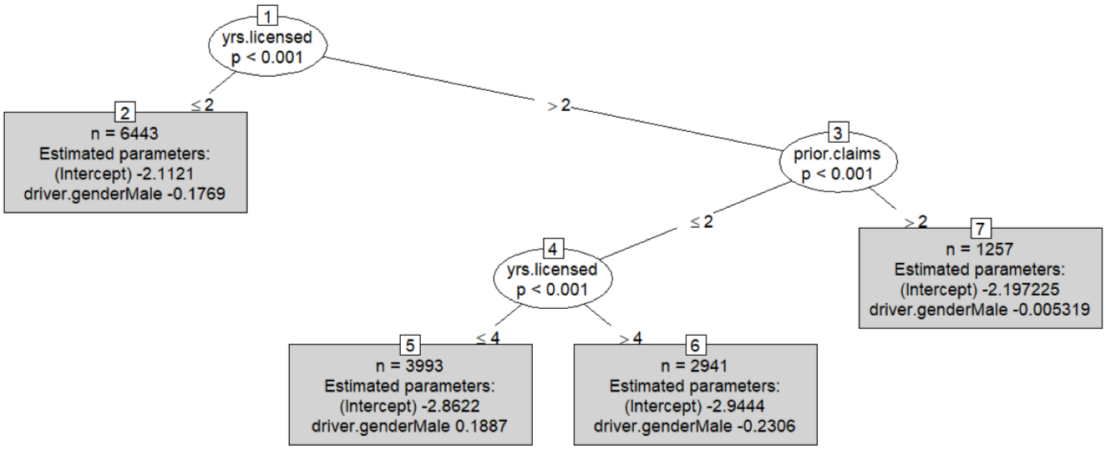
a) (**2*p.***) Omów jedną wybraną regułę określającą, kiedy węzeł w drzewie regresyjnym jest przyjmowany za końcowy (jest uznawany za liść).

b) (**3*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),
driver.gender – płeć kierowcy (zmienna jakościowa przyjmująca dwie kategorie: Female i Male.

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 (grupa A) i grupę, która średnio rocznie zgłasza najmniej szkód (grupa B). Dla grupy A oszacuj prawdopodobieństwa wystąpienia co najmniej jednego roszczenia.

Rys. 10.1



a) Proces podziału drzewa jest zatrzymywany, gdy osiągnie ono zdefiniowaną liczbę poziomów (głębokość). Głębokość węzła jest liczona od korzenia drzewa (który ma głębokość zero), a jego bezpośredni potomkowie mają głębokość o jeden większą.

b) Należy obliczyć oczekiwaną liczbę szkód dla każdej płci dla wszystkich liści.

Z linku kanonicznego w regresji Poissona:

$\lambda = \exp\{\beta_0 + \beta_1 \cdot \text{driver.genderMale}\}$

- Liść 2:

$\lambda_K = \exp\{-2.1121\} = 0.1210$

$\lambda_M = \exp\{-2.1121 - 0.1769\} = 0.1015$

- Liść 5:

$\lambda_K = \exp\{-2.8622\} = 0.0571$

$\lambda_M = \exp\{-2.8622 + 0.1887\} = 0.0690$

- Lit-t 6:

$\lambda_K = \exp\{-2.9444\} = 0.0526$

$\lambda_M = \exp\{-2.9444 - 0.2306\} = 0.0418$

- Lit-t 7:

$\lambda_K = \exp\{-2.197225\} = 0.1111$

$\lambda_M = \exp\{-2.197225 - 0.005319\} = 0.1105$

Grupa A: kobiety, które posiadają prawo jazdy nie dłużej niż 2 lata (niezależnie od liczby zgłoszonych wcześniej szkód i wieku)

$\lambda = 0.1210$

$P(k \geq 1) = 1 - P(k=0) = 1 - e^{-0.121} = 0.114$

Grupa B: mężczyźni, którzy posiadają prawo jazdy powyżej 4 lat i wcześniej zgłosili co najwyżej 2 szkody (niezależnie od wieku)

$\lambda = 0.0418$