Komisja Egzaminacyjna dla Aktuariuszy

XCVI Egzamin dla Aktuariuszy

Sesja egzaminacyjna w dniu 26 maja 2026 r.

Modelowanie

Numer rejestru: _____

Czas trwania egzaminu: **120 minut**

Uwagi

A) W prezentowanych wynikach separatorem dziesiętnym (znakiem dziesiętnym) jest kropka ".".

B) W prezentowanych wynikach oszacowań modeli:

- `Residual deviance` i `Resid. Dev` – oznacza dewiancję oszacowanego modelu,
- `Null deviance` – oznacza dewiancję modelu zerowego,
- `Deviance` – redukcję dewiancji po dodaniu kolejnej zmiennej objaśniającej,
- `Df` – stopnie swobody,
- `Sum Sq` – suma kwadratów,
- `log Lik.` – logarytm wiarygodności,
- `log` – logarytm naturalny.

C) **Dystrybuanta standardowego rozkładu normalnego.**

	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.500	0.504	0.508	0.512	0.516	0.520	0.524	0.528	0.532	0.536
0.1	0.540	0.544	0.548	0.552	0.556	0.560	0.564	0.568	0.571	0.575
0.2	0.579	0.583	0.587	0.591	0.595	0.599	0.603	0.606	0.610	0.614
0.3	0.618	0.622	0.626	0.630	0.634	0.638	0.642	0.645	0.649	0.653
0.4	0.655	0.659	0.663	0.668	0.671	0.675	0.679	0.682	0.686	0.689
0.5	0.692	0.696	0.699	0.702	0.706	0.709	0.712	0.715	0.718	0.722
0.6	0.725	0.728	0.731	0.734	0.737	0.740	0.742	0.745	0.748	0.751
0.7	0.754	0.757	0.760	0.762	0.765	0.768	0.770	0.773	0.776	0.778
0.8	0.781	0.783	0.786	0.788	0.791	0.793	0.796	0.798	0.800	0.802
0.9	0.805	0.807	0.809	0.812	0.814	0.816	0.818	0.820	0.822	0.824
1	0.826	0.828	0.830	0.832	0.834	0.836	0.838	0.840	0.842	0.844
1.1	0.846	0.848	0.850	0.851	0.853	0.855	0.857	0.858	0.860	0.862
1.2	0.864	0.865	0.867	0.869	0.870	0.872	0.874	0.875	0.877	0.878
1.3	0.880	0.881	0.883	0.884	0.886	0.887	0.889	0.890	0.891	0.893
1.4	0.894	0.896	0.897	0.898	0.900	0.901	0.902	0.904	0.905	0.906
1.5	0.908	0.909	0.910	0.911	0.913	0.914	0.915	0.916	0.918	0.919
1.6	0.920	0.921	0.923	0.924	0.925	0.926	0.927	0.928	0.929	0.930
1.7	0.931	0.932	0.933	0.934	0.935	0.936	0.937	0.938	0.939	0.940
1.8	0.941	0.942	0.943	0.944	0.945	0.946	0.946	0.947	0.948	0.949
1.9	0.950	0.950	0.951	0.952	0.953	0.953	0.954	0.955	0.955	0.956
2	0.977	0.958	0.959	0.959	0.960	0.961	0.961	0.962	0.963	0.963
2.1	0.964	0.964	0.965	0.966	0.966	0.967	0.967	0.968	0.968	0.969
2.2	0.970	0.970	0.971	0.971	0.972	0.972	0.973	0.973	0.974	0.974
2.3	0.975	0.975	0.976	0.976	0.977	0.977	0.978	0.978	0.979	0.979
2.4	0.980	0.980	0.980	0.981	0.981	0.982	0.982	0.983	0.983	0.983
2.5	0.984	0.984	0.985	0.985	0.986	0.986	0.986	0.987	0.987	0.988
2.6	0.988	0.989	0.989	0.990	0.990	0.990	0.991	0.991	0.991	0.992
2.7	0.992	0.993	0.993	0.993	0.994	0.994	0.994	0.995	0.995	0.995
2.8	0.997	0.997	0.997	0.997	0.998	0.998	0.998	0.998	0.999	0.999
2.9	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
3	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999

D) **Wartości rozkładu chi-kwadrat $\chi^2_{\nu,\alpha}$ spełniające warunek $\mathbb{P}(\chi^2_\nu > \chi^2_{\nu,\alpha}) = \alpha$.**

	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801

E) **Wartości rozkładu t-Studenta $t_{\nu,\alpha}$ spełniające warunek $\mathbb{P}(T_\nu > t_{\nu,\alpha}) = \alpha$.**

	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	0.008	0.016	0.039	0.079	0.158	3.078	6.314	12.706	25.452	127.32
2	0.033	0.067	0.142	0.289	0.509	1.886	2.920	4.303	6.965	9.925
3	0.055	0.099	0.176	0.277	0.424	1.638	2.353	3.182	4.541	5.841
4	0.073	0.116	0.181	0.238	0.355	1.533	2.132	2.776	3.747	4.604
5	0.087	0.129	0.189	0.246	0.374	1.476	2.015	2.571	3.365	4.032
6	0.099	0.141	0.198	0.256	0.395	1.440	1.943	2.447	3.143	3.707
7	0.108	0.152	0.207	0.267	0.417	1.415	1.895	2.365	2.998	3.499
8	0.117	0.162	0.217	0.280	0.440	1.397	1.860	2.306	2.896	3.355
9	0.125	0.172	0.227	0.292	0.463	1.383	1.833	2.262	2.821	3.250
10	0.132	0.180	0.237	0.306	0.487	1.372	1.812	2.228	2.764	3.169
20	0.164	0.214	0.277	0.352	0.495	1.325	1.725	2.086	2.528	2.845
30	0.184	0.235	0.303	0.387	0.579	1.310	1.697	2.042	2.457	2.750
60	0.210	0.267	0.349	0.448	0.673	1.296	1.671	2.000	2.390	2.660
120	0.222	0.282	0.369	0.477	0.717	1.289	1.658	1.980	2.358	2.617
500	0.232	0.295	0.387	0.504	0.763	1.283	1.648	1.965	2.345	2.599
1000	0.006	0.013	0.031	0.063	0.126	1.646	1.962	2.245	2.581	2.813

Zadanie 1

W pewnym portfelu ubezpieczeń komunikacyjnych modelowano roczną liczbę szkód N_i za pomocą regresji Poissona. Dla każdej polisy znane są:

$$Exposure_i \quad \text{– ekspozycja w latach,}$$

$$DriverAge_i \quad \text{– wiek kierowcy,}$$

$$VehPower_{c,i} = VehPower_i - 6 \quad \text{– scentrowana moc pojazdu,}$$

oraz zmienna jakościowa $Area_i \in \{A, B, C\}$, gdzie kategorią referencyjną jest A.

Przyjęto model regresji Poissona z logarytmiczną funkcją łączącą oraz offsetem $\log(Exposure_i)$. W modelu uwzględniono efekty główne zmiennych $DriverAge$, $VehPower_c$, $Area$ oraz interakcje między scentrowaną mocą pojazdu $VehPower_c$ a obszarem użytkowania pojazdu $Area$.

Model ma postać:

$$N_i \sim \text{Poisson}(\mu_i),$$

$$\begin{aligned}\log(\mu_i) = \log(Exposure_i) &+ \beta_0 + \beta_1 DriverAge_i + \beta_2 VehPower_{c,i} \\ &+ \beta_3 AreaB_i + \beta_4 AreaC_i \\ &+ \beta_5 VehPower_{c,i} : AreaB_i + \beta_6 VehPower_{c,i} : AreaC_i.\end{aligned}$$

Otrzymano następujące oszacowania parametrów:

Zmienna	Estimate
Intercept	-2.7500
DriverAge	-0.0110
VehPower_c	0.0640
AreaB	0.2450
AreaC	0.5100
VehPower_c:AreaB	-0.0180
VehPower_c:AreaC	-0.0300

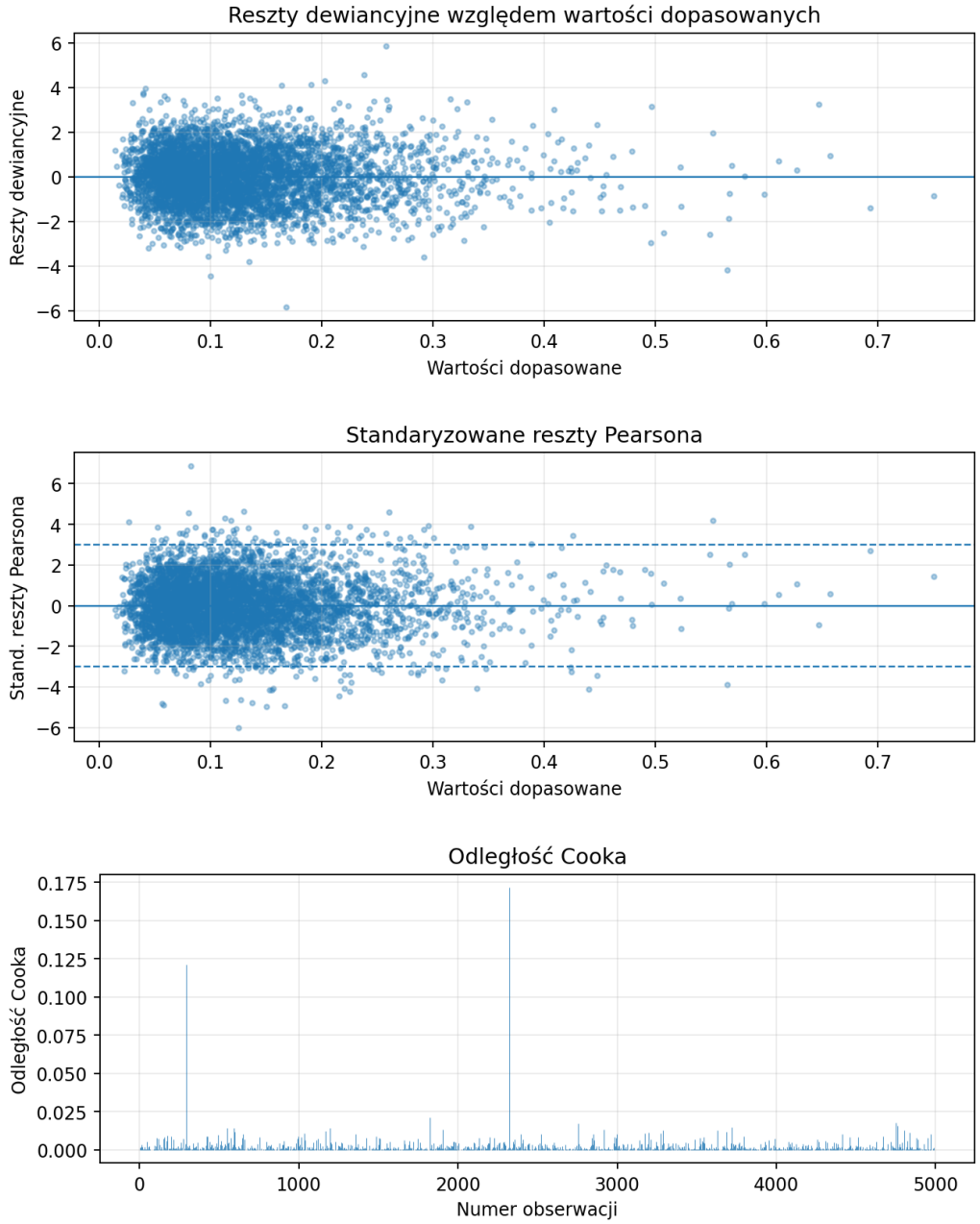
Dla modelu suma kwadratów reszt Pearsona wyniosła:

$$\sum_{i=1}^{n} \left(r_i^{(P)}\right)^2 = 7215.9,$$

a liczba stopni swobody reszt była równa:

$$df = 4993.$$

Na rysunku 1.1 przedstawiono wykresy diagnostyczne: (i) reszty dewiancyjne względem wartości dopasowanych, (ii) standaryzowane reszty Pearsona względem wartości dopasowanych, (iii) odległość Cooka.



Rysunek 1.1: Wykresy diagnostyczne dla modelu regresji Poissona.

a) (2 pkt) Dla polisy o parametrach:

$$Exposure = 0.5, \qquad DriverAge = 40, \qquad VehPower = 7, \qquad Area = C,$$

podaj wiersz macierzy modelu X, przyjmując kolejność kolumn jak w tabeli współczynników. Następnie oblicz prognozowaną wartość $E[N]$ dla tej polisy.

b) (1 pkt) Oblicz estymator parametru dyspersji i zinterpretuj wynik. Wyjaśnij, jak przejście z modelu Poissona do modelu quasi-Poissona wpłynęłoby na błędy standardowe ocen parametrów, tj. na $SE(\hat{\beta}_j)$.

c) (2 pkt) Na podstawie rysunku 1.1 wskaż dwa potencjalne problemy diagnostyczne modelu. W odpowiedzi odnieś się do co najmniej dwóch z trzech wykresów oraz zaproponuj jedną możliwą modyfikację modelu.

Rozwiązanie

Zakres literatury

Tematyka zadania jest omawiana w szczególności w następujących pozycjach z wykazu literatury:

- M. Denuit, D. Hainaut, J. Trufin, *Effective Statistical Learning Methods for Actuaries I*, Springer, 2019: rozdziały 4–5.2, w szczególności: GLM, dewiancja, estymacja, zmienne offset, ocena dopasowania, testy statystyczne, model Poissona i model ujemny dwumianowy; rozdziały 7.3–7.4: modelowanie współczynnika dyspersji w modelach GLM.
- M.V. Wüthrich, M. Merz, *Statistical Foundations of Actuarial Learning and its Applications*, rozdział 5: GLM, dewiancja, offset, ocena dopasowania, model Poissona, model ujemny dwumianowy oraz modelowanie współczynnika dyspersji.
- E.W. Frees, *Regression Modeling with Actuarial and Financial Applications*, Cambridge, 2009: rozdział 12 – regresja Poissona oraz rozdział 13 – uogólnione modele liniowe.

a) Dla rozważanej polisy:

$$Exposure = 0.5, \qquad DriverAge = 40, \qquad VehPower = 7, \qquad Area = C.$$

Ponieważ:

$$VehPower_c = VehPower - 6 = 1,$$

oraz dla kategorii C:

$$AreaB = 0, \qquad AreaC = 1,$$

otrzymujemy:

$$VehPower_c : AreaB = 0, \qquad VehPower_c : AreaC = 1.$$

Przy kolejności kolumn jak w tabeli współczynników:

$(Intercept)$, $DriverAge$, $VehPower_c$, $AreaB$, $AreaC$, $VehPower_c : AreaB$, $VehPower_c : AreaC$,

wiersz macierzy modelu ma postać:

$$\boxed{x_i = (1,\ 40,\ 1,\ 0,\ 1,\ 0,\ 1).}$$

Predyktor liniowy z offsetem wynosi:

$$\hat{\eta}_i = \log(0.5) - 2.7500 - 0.0110 \cdot 40 + 0.0640 + 0.5100 - 0.0300 \approx -3.3391.$$

Zatem:

$$\widehat{E[N_i]} = \hat{\mu}_i = \exp(\hat{\eta}_i) \approx \exp(-3.3391) \approx 0.0355.$$

Ostatecznie:

$$\boxed{\widehat{E[N_i]} \approx 0.036.}$$

Model przewiduje więc około 0.036 szkody dla tej polisy przy ekspozycji 0.5 roku.

b) Estymator parametru dyspersji oparty na resztach Pearsona wynosi:

$$\hat{\phi} = \frac{\sum_{i=1}^{n} \left(r_i^{(P)}\right)^2}{df} = \frac{7215.9}{4993} \approx 1.45.$$

Zatem:

$$\boxed{\hat{\phi} \approx 1.45.}$$

W klasycznym modelu Poissona zakłada się:

$$\mathrm{Var}(N_i) = \mu_i,$$

czyli $\phi = 1$. Otrzymana wartość $\hat{\phi} > 1$ wskazuje na nadmierną dyspersję, a więc zmienność liczby szkód większą niż wynikałoby to z modelu Poissona.

W modelu quasi-Poissona przyjmuje się:

$$\mathrm{Var}(N_i) = \phi\mu_i.$$

Przejście do modelu quasi-Poissona zwykle nie zmienia ocen parametrów $\hat{\beta}_j$, ale zwiększa błędy standardowe ocen parametrów:

$$SE(\hat{\beta}_j) \quad \text{mnoży się w przybliżeniu przez} \quad \sqrt{\hat{\phi}} \approx \sqrt{1.45} \approx 1.20.$$

Błędy standardowe byłyby więc około 20% większe, co obniżałoby wartości statystyk testowych i zwiększałoby wartości p-value.

c) Na podstawie wykresów diagnostycznych można wskazać następujące potencjalne problemy modelu.

Po pierwsze, wykres standaryzowanych reszt Pearsona pokazuje obserwacje przekraczające poziomy około

$$-3 \quad \text{oraz} \quad 3.$$

Oznacza to występowanie obserwacji potencjalnie odstających. Jest to zgodne z wynikiem z punktu b), czyli z sygnałem nadmiernej dyspersji.

Po drugie, na wykresie reszt dewiancyjnych względem wartości dopasowanych widać rozrzut reszt oraz pojedyncze obserwacje wyraźnie oddalone od zera, zwłaszcza dla większych wartości dopasowanych. Może to oznaczać, że model nie opisuje w pełni zmienności w portfelu, szczególnie dla polis o wyższym przewidywanym ryzyku.

Po trzecie, wykres odległości Cooka wskazuje kilka obserwacji o wyraźnie większym wpływie niż pozostałe. Takie obserwacje należy przeanalizować, ponieważ mogą istotnie wpływać na oszacowania parametrów.

Możliwa modyfikacja modelu:

$$\boxed{\text{zastosowanie modelu quasi-Poissona albo modelu ujemnego dwumianowego.}}$$

Alternatywnie można rozważyć dodanie nieliniowych efektów wybranych zmiennych, np. splajnów dla wieku kierowcy lub mocy pojazdu, dodatkowych interakcji albo osobną analizę obser-

wacji wpływowych.

Wniosek:

> Diagnostyka wskazuje przede wszystkim na nadmierną dyspersję oraz obserwacje odstające lub wpływowe.

Zadanie 2

W pewnym zakładzie ubezpieczeń analizowano rezygnację z odnowienia polisy komunikacyjnej. Niech $Y_i = 1$ oznacza rezygnację z odnowienia polisy, $Y_i = 0$ oznacza odnowienie polisy, a $\pi_i = P(Y_i = 1)$.

W modelu wykorzystano następujące zmienne objaśniające:

- $Age_{c,i} = Age_i - 40$, gdzie Age_i oznacza wiek klienta w latach; zmienna jest scentrowana wokół 40 lat,
- $Tenure_i$ – staż klienta w zakładzie ubezpieczeń, wyrażony w latach,
- $ChannelOnline_i$ – zmienna zero-jedynkowa przyjmująca wartość 1 dla klientów obsługiwanych przez kanał online oraz 0 dla klientów obsługiwanych przez agenta,
- $SegmentHigh_i$ – zmienna zero-jedynkowa przyjmująca wartość 1 dla klientów z segmentu wysokiej składki oraz 0 dla pozostałych klientów.

Rozważono modele logitowe:

$$M_1: \quad \log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 Age_{c,i} + \beta_2 Tenure_i + \beta_3 ChannelOnline_i + \beta_4 SegmentHigh_i,$$

oraz model M_2, który dodatkowo zawiera interakcje:

$$Age_c : ChannelOnline \quad \text{oraz} \quad Tenure : SegmentHigh.$$

Dla modelu M_2 otrzymano:

Zmienna	Estimate	Std. Error	z value	p-value
Intercept	-1.850	0.090	-20.556	$< 2 \cdot 10^{-16}$
Age_c	-0.030	0.006	-5.000	$5.7 \cdot 10^{-7}$
Tenure	-0.120	0.020	-6.000	$2.0 \cdot 10^{-9}$
ChannelOnline	0.420	0.110	3.818	0.0001
SegmentHigh	0.760	0.130	5.846	$5.0 \cdot 10^{-9}$
Age_c:ChannelOnline	0.022	0.008	2.750	0.0060
Tenure:SegmentHigh	-0.080	0.030	-2.667	0.0077

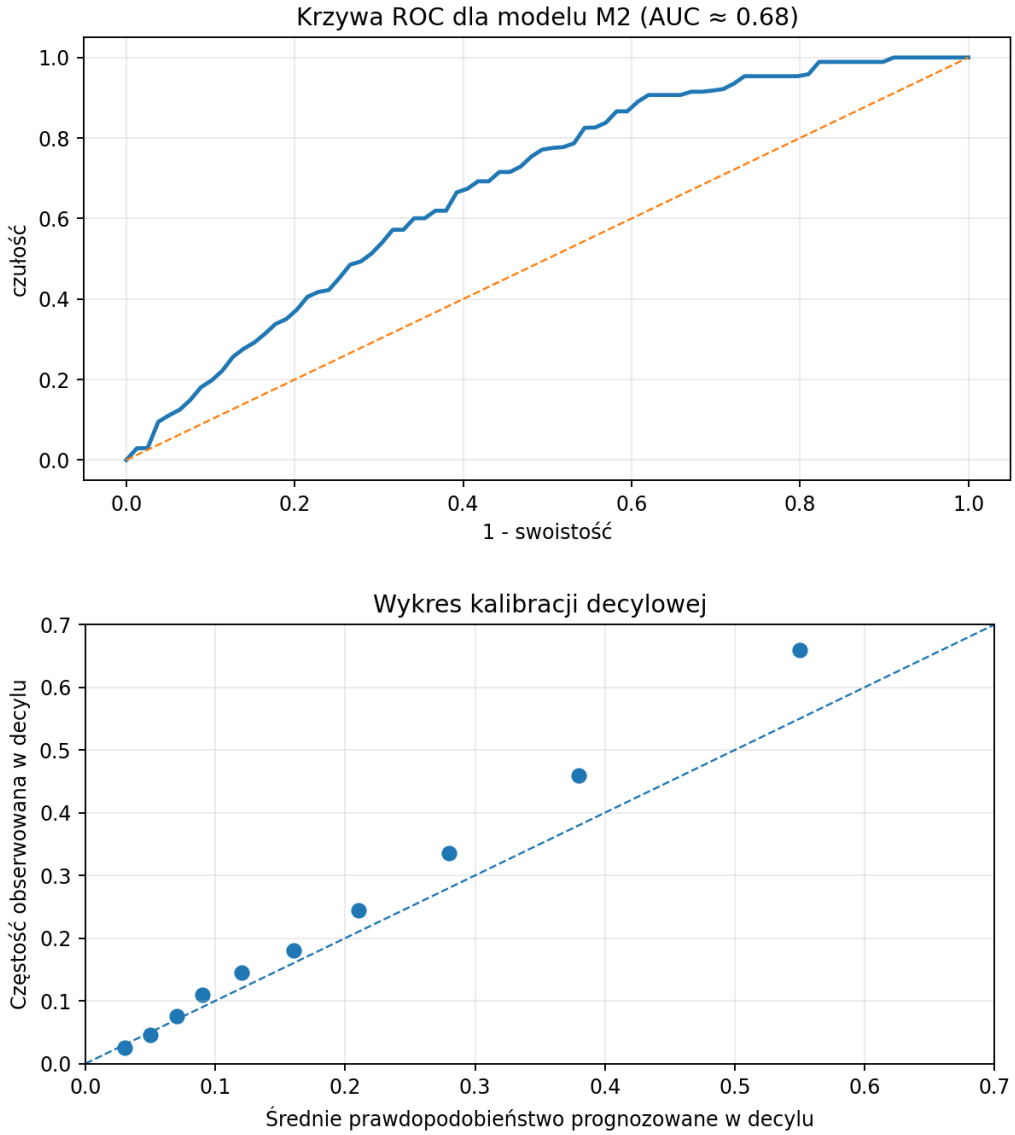
Ponadto:

$$D(M_2) = 1831.4, \quad df(M_2) = 2993, \quad AIC(M_2) = 1845.4,$$

$$D(M_1) = 1842.6, \quad df(M_1) = 2995, \quad AIC(M_1) = 1852.6.$$

W powyższym zapisie $D(M)$ oznacza dewiancję resztową modelu M, a $df(M)$ oznacza liczbę stopni swobody reszt.

Na rysunku 2.1 przedstawiono krzywą ROC oraz wykres kalibracji decylowej dla modelu M_2.



Rysunek 2.1: Krzywa ROC i wykres kalibracji decylowej dla modelu logitowego M_2.

a) (2 pkt) Zinterpretuj interakcję Age_c : $ChannelOnline$. Oblicz iloraz szans rezygnacji dla kanału online względem kanału agenta dla klienta w wieku 40 lat oraz dla klienta w wieku 55 lat, przy pozostałych zmiennych ustalonych. Wyjaśnij, czy wpływ kanału online na rezygnację rośnie, czy maleje wraz z wiekiem klienta.

b) (1 pkt) Zweryfikuj na poziomie istotności 0.05, czy dodanie interakcji w modelu M_2 istotnie poprawia dopasowanie względem modelu M_1.

c) (2 pkt) Na podstawie rysunku 2.1 oceń jakość modelu M_2 z punktu widzenia: (i) zdolności porządkowania ryzyk, (ii) kalibracji prognozowanych prawdopodobieństw. Wskaż jedną możliwą modyfikację modelu albo procedury wdrożeniowej, jeśli model miałby być wykorzystywany do decyzji taryfowych lub retencyjnych.

Rozwiązanie

Zakres literatury

Tematyka zadania jest omawiana w szczególności w następujących pozycjach z wykazu literatury:

- M. Denuit, D. Hainaut, J. Trufin, *Effective Statistical Learning Methods for Actuaries I*, Springer, 2019: rozdziały 4–5.2, w szczególności uogólnione modele liniowe, dewiancja, estymacja, ocena dopasowania modelu i testy statystyczne.
- M.V. Wüthrich, M. Merz, *Statistical Foundations of Actuarial Learning and its Applications*, rozdział 4.1–4.2: ocena dopasowania i zdolności predykcyjnej modeli, w tym kryterium AIC, oraz rozdział 5: GLM, dewiancja, estymacja i testy statystyczne.
- G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, 2021: rozdziały 4.1–4.5 i 4.7.2, w szczególności klasyfikacja i regresja logistyczna.
- E.W. Frees, *Regression Modeling with Actuarial and Financial Applications*, Cambridge, 2009: rozdział 11 – regresja logistyczna.

a) W modelu M_2 wpływ kanału online na logarytm ilorazu szans rezygnacji zależy od wieku klienta. Dla ustalonej wartości Age_c, przy pozostałych zmiennych bez zmian, logarytm ilorazu szans rezygnacji dla kanału online względem kanału agenta wynosi:

$$\log(OR_{\text{online/agent}}) = \hat{\beta}_{ChannelOnline} + \hat{\beta}_{Age_c:ChannelOnline} Age_c.$$

Dla klienta w wieku 40 lat mamy $Age_c = 0$, więc:

$$OR_{\text{online/agent}}(40) = \exp(0.420) \approx 1.52.$$

Dla klienta w wieku 55 lat mamy $Age_c = 15$, więc:

$$OR_{\text{online/agent}}(55) = \exp(0.420 + 0.022 \cdot 15) = \exp(0.750) \approx 2.12.$$

Zatem:

$$\boxed{OR_{\text{online/agent}}(40) \approx 1.52, \qquad OR_{\text{online/agent}}(55) \approx 2.12.}$$

Dodatnia wartość współczynnika przy interakcji

$$\widehat{\beta}_{Age_c:ChannelOnline} = 0.022$$

oznacza, że wpływ kanału online na szanse rezygnacji rośnie wraz z wiekiem klienta. Innymi słowy, różnica między klientami obsługiwanymi online i przez agenta jest większa dla starszych klientów.

b) Testujemy hipotezy:

$$H_0: \ \beta_{Age_c:ChannelOnline} = 0, \quad \beta_{Tenure:SegmentHigh} = 0,$$

$$H_1: \ \text{co najmniej jeden z powyższych współczynników jest różny od zera.}$$

Statystyka testu ilorazu wiarygodności dla modeli zagnieżdżonych może być zapisana jako różnica dewiancji:

$$D = D(M_1) - D(M_2) = 1842.6 - 1831.4 = 11.2.$$

Różnica liczby stopni swobody wynosi:

$$df(M_1) - df(M_2) = 2995 - 2993 = 2.$$

Przy H_0:

$$D \overset{a}{\sim} \chi^2(2).$$

Na poziomie istotności 0.05 wartość krytyczna wynosi około:

$$\chi^2_{2;0.05} = 5.99.$$

Ponieważ:

$$11.2 > \chi^2_{2;0.05},$$

odrzucamy hipotezę zerową. Dodanie interakcji w modelu M_2 istotnie poprawia dopasowanie względem modelu M_1.

Wniosek jest zgodny z kryterium AIC:

$$AIC(M_2) = 1845.4 < 1852.6 = AIC(M_1).$$

Niższa wartość AIC również przemawia za wyborem modelu M_2.

c) Krzywa ROC dla modelu M_2 leży powyżej linii losowej klasyfikacji, a wartość

$$AUC \approx 0.68$$

wskazuje na umiarkowaną zdolność porządkowania ryzyk. Model jest lepszy od losowego porządkowania, ale jego moc dyskryminacyjna nie jest bardzo wysoka.

Wykres kalibracji decylowej wskazuje, że dla wielu decyli, zwłaszcza wyższych, częstość obserwowana jest większa od średniego prognozowanego prawdopodobieństwa. Oznacza to, że model może zaniżać prawdopodobieństwo rezygnacji w grupach o wyższym prognozowanym ryzyku.

Możliwa modyfikacja modelu albo procedury wdrożeniowej:

przeprowadzić kalibrację prognozowanych prawdopodobieństw albo rozszerzyć model o efekty nieliniowe.

Zadanie 3

W pewnym portfelu ubezpieczeń majątkowych analizowano duże szkody X_i wyrażone w tys. zł. Dostępna jest próba licząca $n = 5000$ szkód. Aktuariusz chce oszacować wysoki kwantyl rozkładu szkód z wykorzystaniem metody POT (*Peaks Over Threshold*).

Dla ustalonego progu u rozważane są nadwyżki:

$$Y_i = X_i - u \mid X_i > u.$$

Przyjęto, że dla odpowiednio wysokiego progu u rozkład nadwyżek można przybliżyć uogólnionym rozkładem Pareto (GPD). Funkcja przeżycia dla Y_i ma postać:

$$P(Y > y \mid X > u) = \left(1 + \xi \frac{y}{\beta}\right)^{-1/\xi}, \qquad \xi \neq 0, \quad \beta > 0.$$

Prawdopodobieństwo przekroczenia progu u estymujemy empirycznie:

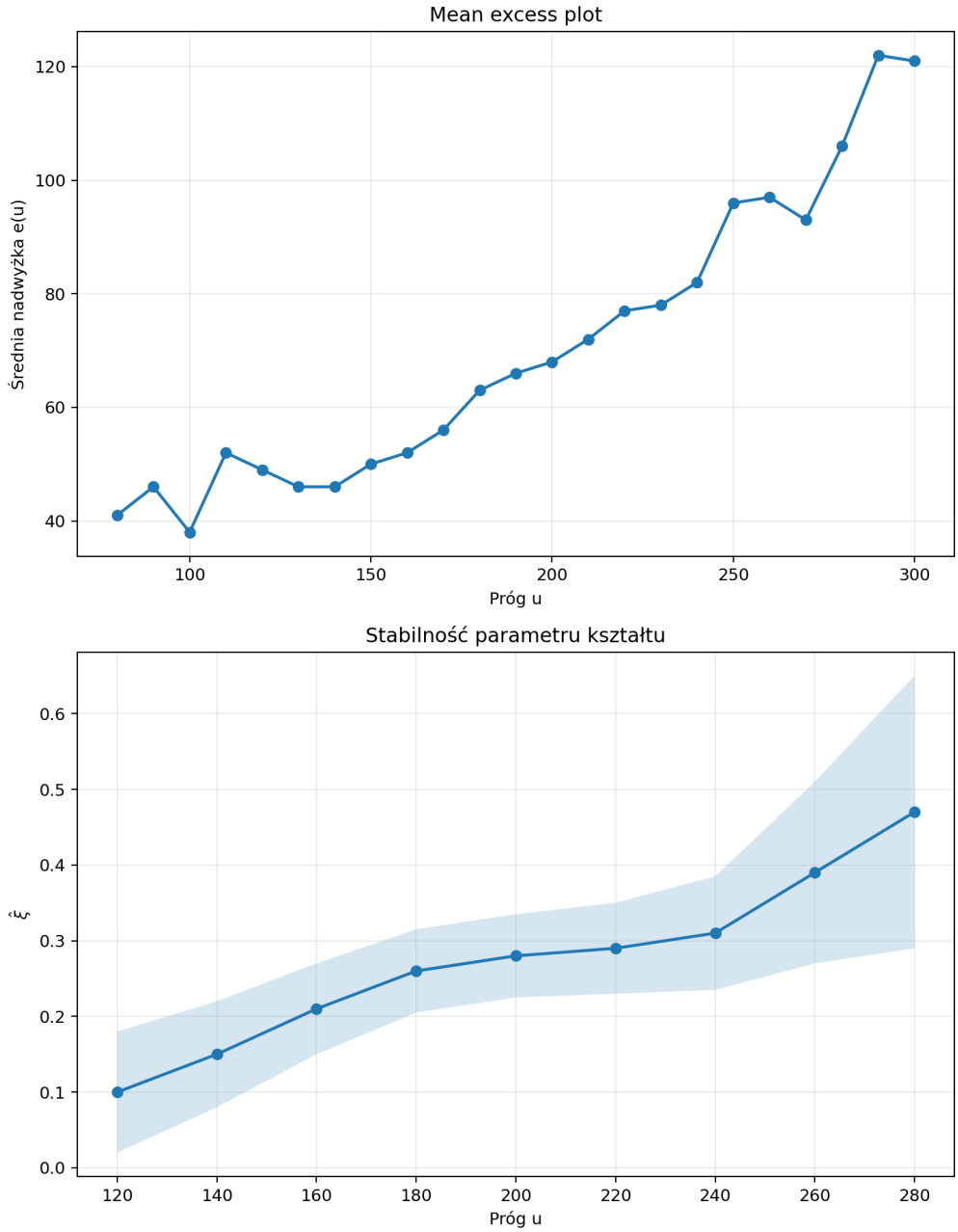
$$P(X > u) \approx \frac{N_u}{n},$$

gdzie N_u oznacza liczbę obserwacji większych od progu u.

Dla wybranych progów otrzymano następujące oszacowania parametrów GPD:

Próg u	Liczba przekroczeń N_u	$\hat{\xi}$	$\widehat{SE}(\hat{\xi})$	$\hat{\beta}$
150	540	0.18	0.045	70
175	410	0.24	0.050	76
200	310	0.28	0.055	85
225	210	0.29	0.065	96
250	120	0.41	0.110	122

Na rysunku 3.1 przedstawiono wykresy diagnostyczne: (i) mean excess plot, (ii) wykres stabilności parametru kształtu ξ.



Rysunek 3.1: Wykresy diagnostyczne służące do wyboru progu oraz oceny dopasowania modelu POT.

a) (2 pkt) Na podstawie tabeli oraz rysunku 3.1 wskaż, który próg u (ze wskazanych w tej tabeli) należy uznać za najbardziej uzasadniony do modelu POT. W odpowiedzi odnieś się do stabilności parametru $\hat{\xi}$, kształtu mean excess plot oraz liczby przekroczeń.

b) (3 pkt) Dla wybranego progu wyprowadź postać estymatora funkcji przeżycia szkody X dla $x > u$, korzystając z zależności:

$$P(X > x) = P(X > u)\, P(X > x \mid X > u).$$

Następnie oszacuj kwantyl rzędu 0.995 dla szkody X.

Rozwiązanie

Zakres literatury

Tematyka zadania jest omawiana w szczególności w następujących pozycjach z wykazu literatury:

- M. Denuit, D. Hainaut, J. Trufin, *Effective Statistical Learning Methods for Actuaries I*, Springer, 2019: rozdział 9 – teoria wartości ekstremalnych, wykresy kwantylowe, wykresy dalszego trwania życia, rozkład nadwyżki szkody, miary ryzyka w ogonie oraz model POT.
- A.J. McNeil, R. Frey, P. Embrechts, *Quantitative Risk Management: Concepts, Techniques and Tools*, Princeton, 2015: rozdział 5 – teoria wartości ekstremalnych, rozkład nadwyżek, model POT oraz estymacja wysokich kwantyli i miar ryzyka w ogonie.
- H. Albrecher, J. Beirlant, J. Teugels, *Reinsurance: Actuarial and Statistical Aspects*, Wiley, 2017: rozdziały 3.4, 4.1 oraz 3.5, 4.2 – wykresy diagnostyczne dla ogona, modele dużych szkód i teoria wartości ekstremalnych.

a) Najbardziej uzasadnionym wyborem progu jest:

$$\boxed{u = 200.}$$

Dla niższych progów pozostaje więcej przekroczeń, ale oszacowania parametru kształtu są jeszcze wyraźnie zmienne i mogą obejmować obserwacje spoza właściwego ogona rozkładu. Dla progów $u = 200$ oraz $u = 225$ wartości

$$\hat{\xi}_{200} = 0.28, \qquad \hat{\xi}_{225} = 0.29$$

są już stabilne. Jednocześnie dla progu $u = 200$ pozostaje więcej przekroczeń:

$$N_{200} = 310,$$

podczas gdy dla $u = 225$ tylko

$$N_{225} = 210.$$

Mean excess plot od okolic $u = 180$–200 ma rosnący, w przybliżeniu liniowy charakter, co jest zgodne z ciężkim ogonem i dodatnią wartością parametru ξ. Próg $u = 250$ wydaje się mniej korzystny, ponieważ liczba przekroczeń spada do 120, a oszacowanie $\hat{\xi} = 0.41$ ma znacznie większy błąd standardowy.

Wybór $u = 200$ jest więc kompromisem między obciążeniem wynikającym ze zbyt niskiego progu a dużą wariancją estymacji przy zbyt wysokim progu.

b) Dla $x > u$:

$$P(X > x) = P(X > u)\, P(X > x \mid X > u).$$

Ponieważ $Y = X - u$, mamy:

$$P(X > x \mid X > u) = P(Y > x - u \mid X > u).$$

Korzystając z funkcji przeżycia GPD oraz z empirycznego oszacowania prawdopodobieństwa

przekroczenia progu:

$$P(X > u) \approx \frac{N_u}{n},$$

otrzymujemy estymator funkcji przeżycia szkody X dla $x > u$:

$$\boxed{\widehat{P}(X > x) = \frac{N_u}{n}\left(1 + \hat{\xi}\frac{x-u}{\hat{\beta}}\right)^{-1/\hat{\xi}}}.$$

Dla wybranego progu $u = 200$:

$$n = 5000, \qquad N_u = 310, \qquad \hat{\xi} = 0.28, \qquad \hat{\beta} = 85.$$

Kwantyl rzędu p spełnia:

$$P(X > q_p) = 1 - p.$$

Po przekształceniu otrzymujemy:

$$\boxed{\hat{q}_p = u + \frac{\hat{\beta}}{\hat{\xi}}\left[\left(\frac{n}{N_u}(1-p)\right)^{-\hat{\xi}} - 1\right]}.$$

Dla $p = 0.995$:

$$\hat{q}_{0.995} = 200 + \frac{85}{0.28}\left[\left(\frac{5000}{310}\cdot 0.005\right)^{-0.28} - 1\right] \approx 510.8.$$

Zatem:

$$\boxed{\hat{q}_{0.995} \approx 511}$$

tys. zł.

Interpretacja: według modelu POT około 0.5% szkód przekracza poziom około 511 tys. zł.

Zadanie 4

W portfelu ubezpieczeń majątkowych analizowano wysokości szkód $X_1, \ldots, X_n$, wyrażone w tys. zł. Dostępna jest próba $n = 300$ niezależnych szkód. Aktuariusz rozważa dwie wielkości:

$$\theta_1 = E[X], \qquad \theta_2 = E[(X - 100)_+],$$

gdzie θ_2 oznacza oczekiwaną nadwyżkę szkody ponad udział własny 100 tys. zł.

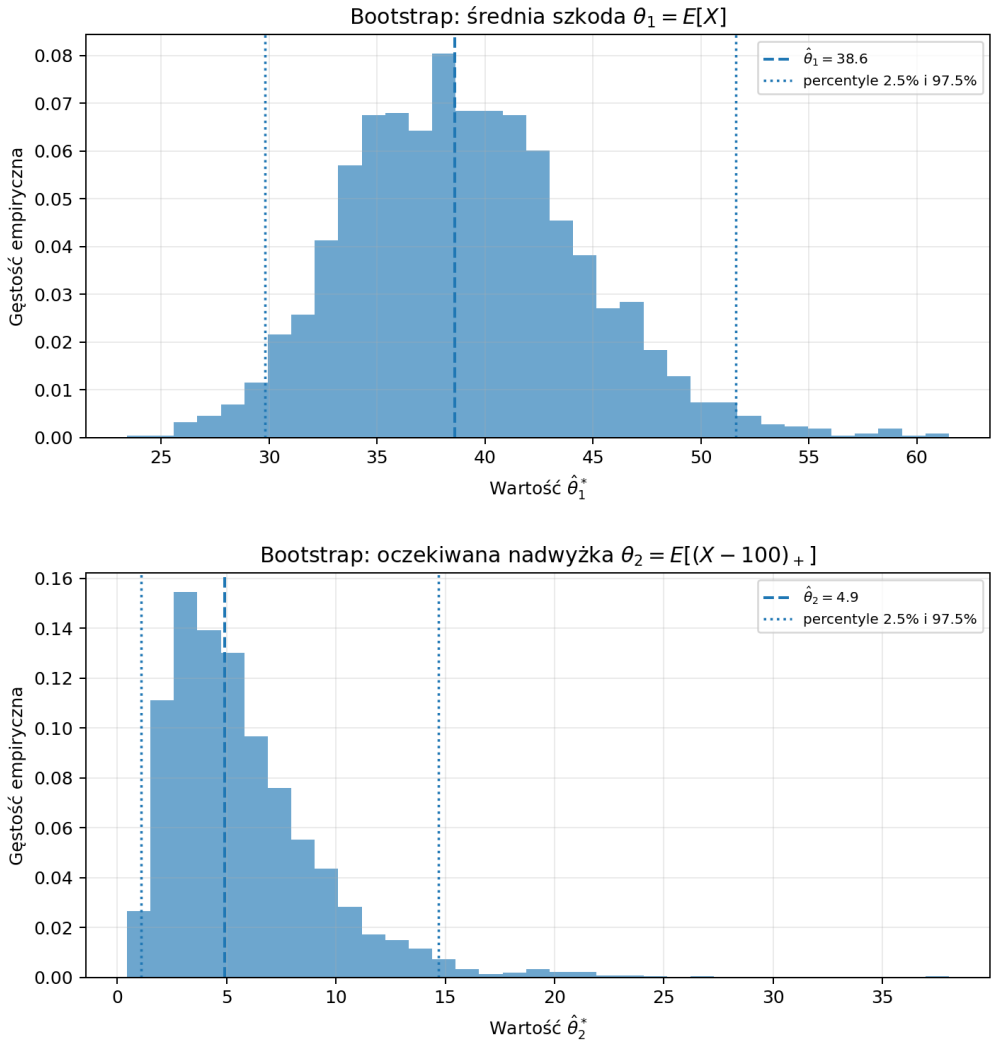
Z próby otrzymano:

$$\hat{\theta}_1 = \bar{X} = 38.6, \qquad \hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - 100)_+ = 4.9.$$

Wykonano $B = 2000$ replikacji bootstrapu nieparametrycznego. Otrzymano następujące wyniki:

Estymator	Wartość z próby	Średnia bootstrapowa	Odch. stand.	$q_{2.5\%}$	$q_{97.5\%}$
$\hat{\theta}_1^*$	38.6	39.1	5.4	29.8	51.6
$\hat{\theta}_2^*$	4.9	5.8	3.6	1.1	14.7

Na rysunku 4.1 przedstawiono rozkłady bootstrapowe estymatorów $\hat{\theta}_1^*$ oraz $\hat{\theta}_2^*$.



Rysunek 4.1: Rozkłady bootstrapowe estymatorów średniej szkody i oczekiwanej nadwyżki ponad udział własny.

a) (2 pkt) Dla obu estymatorów oblicz obciążenie bootstrapowe. Wyjaśnij, dlaczego estymator $\theta_2 = E[(X - 100)_+]$ może być bardziej wrażliwy na pojedyncze duże szkody niż estymator średniej szkody.

b) (1 pkt) Dla obu wielkości podaj 95% przedziały ufności metodą percentylową. Porównaj szerokość względną tych przedziałów, rozumianą jako:

$$\frac{q_{97.5\%} - q_{2.5\%}}{\hat{\theta}}.$$

Która wielkość jest oszacowana z większą względną niepewnością? Zinterpretuj wynik z punktu widzenia wyceny ochrony z udziałem własnym.

c) (2 pkt) Wyjaśnij, dlaczego zwykły bootstrap nieparametryczny może zaniżać niepewność estymacji wielkości ogonowych, jeżeli w próbie znajduje się mało bardzo dużych szkód. Wskaż jedną możliwą alternatywę lub modyfikację metody bootstrapowej w takiej sytuacji.

Rozwiązanie

Zakres literatury

Tematyka zadania jest omawiana w szczególności w następujących pozycjach z wykazu literatury:

- M. Denuit, D. Hainaut, J. Trufin, *Effective Statistical Learning Methods for Actuaries I*, Springer, 2019: rozdział 3.7 – bootstrap.
- M.V. Wüthrich, M. Merz, *Statistical Foundations of Actuarial Learning and its Applications*, rozdział 4.3 – bootstrap.
- M.V. Wüthrich, C. Buser, *Data Analytics for Non-Life Insurance Pricing*, rozdział 7.1 – bootstrap.
- G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, rozdziały 5.2 i 5.3 – bootstrap oraz metody resamplingowe.

a) Obciążenie bootstrapowe estymatora rozumiemy jako różnicę między średnią bootstrapową a wartością estymatora obliczoną z próby.

Dla średniej szkody:

$$\widehat{bias}_{boot}(\hat{\theta}_1) = 39.1 - 38.6 = 0.5.$$

Dla oczekiwanej nadwyżki ponad udział własny:

$$\widehat{bias}_{boot}(\hat{\theta}_2) = 5.8 - 4.9 = 0.9.$$

Zatem:

$$\boxed{\widehat{bias}_{boot}(\hat{\theta}_1) = 0.5, \qquad \widehat{bias}_{boot}(\hat{\theta}_2) = 0.9.}$$

Estymator

$$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - 100)_+$$

jest bardziej wrażliwy na pojedyncze duże szkody niż estymator średniej szkody, ponieważ obserwacje poniżej progu 100 tys. zł nie wnoszą do niego dodatniego wkładu, natomiast pojedyncza bardzo duża szkoda może wygenerować dużą nadwyżkę. W rezultacie wartość $\hat{\theta}_2$ zależy silniej od prawego ogona rozkładu szkód niż zwykła średnia.

b) Percentylowe przedziały ufności 95% odczytujemy bezpośrednio z kwantyli rozkładów bootstrapowych.

Dla średniej szkody:

$$\boxed{CI_{0.95}(\theta_1) = [29.8,\ 51.6].}$$

Dla oczekiwanej nadwyżki:

$$\boxed{CI_{0.95}(\theta_2) = [1.1,\ 14.7].}$$

Względna szerokość przedziału dla θ_1 wynosi:

$$\frac{51.6 - 29.8}{38.6} \approx 0.565.$$

Względna szerokość przedziału dla θ_2 wynosi:

$$\frac{14.7 - 1.1}{4.9} \approx 2.78.$$

Zatem:

θ_2 jest oszacowana ze znacznie większą względną niepewnością.

Interpretacja aktuarialna jest taka, że wycena ochrony ponad udział własny jest dużo bardziej niepewna niż estymacja przeciętnej wysokości szkody. Wynika to z faktu, że składka za ochronę nadwyżkową zależy głównie od rzadkich, dużych szkód, czyli od ogona rozkładu.

c) Zwykły bootstrap nieparametryczny losuje obserwacje ze zwracaniem wyłącznie z danych empirycznych. Oznacza to, że w próbach bootstrapowych mogą pojawić się tylko szkody już zaobserwowane w próbie.

Jeżeli w próbie jest mało bardzo dużych szkód, bootstrap może zaniżać niepewność estymacji wielkości ogonowych. Po pierwsze, nie generuje szkód większych niż największa szkoda zaobserwowana w próbie. Po drugie, rozkład bootstrapowy wielkości ogonowej może być zdominowany przez to, ile razy wylosowano kilka największych obserwacji. W konsekwencji metoda może nie odzwierciedlać pełnej niepewności związanej z prawym ogonem rozkładu.

Możliwą modyfikacją jest zastosowanie bootstrapu parametrycznego lub semiparametrycznego dla ogona rozkładu, np. po dopasowaniu modelu POT/GPD do dużych szkód. W podejściu semiparametrycznym środek rozkładu można próbkować empirycznie, a ogon generować z dopasowanego modelu parametrycznego. Alternatywnie można wykonać analizę wrażliwości na największe szkody albo wykorzystać dłuższy okres obserwacji, jeżeli portfel pozostaje porównywalny.

Dla wielkości ogonowych zwykły bootstrap empiryczny może być niewystarczający.

Zadanie 5

Dla portfela aktywów stanowiących pokrycie rezerw techniczno-ubezpieczeniowych analizowano dzienne logarytmiczne stopy zwrotu:

$$r_t = \log\left(\frac{P_t}{P_{t-1}}\right), \qquad t = 1, \ldots, 1500.$$

Aktuariusz chce ocenić, czy szereg r_t można traktować jako biały szum o stałej wariancji, czy też należy uwzględnić zmienność warunkową.

Dla stóp zwrotu oraz kwadratów stóp zwrotu wykonał testy Ljunga–Boxa, otrzymując następujące wyniki:

Test	Statystyka	df	p-value
Ljung–Box dla r_t, lag = 20	24.8	20	0.211
Ljung–Box dla r_t^2, lag = 20	96.4	20	< 0.001

Następnie oszacował model AR(1)–GARCH(1,1):

$$r_t = \mu + \phi r_{t-1} + a_t, \qquad a_t = \sigma_t z_t, \qquad z_t \sim iid\ N(0,1),$$

$$\sigma_t^2 = \omega + \alpha a_{t-1}^2 + \beta \sigma_{t-1}^2, \qquad \omega > 0, \quad \alpha, \beta \geq 0.$$

Otrzymał następujące oszacowania parametrów:

$$\hat{\mu} = 0.00012, \qquad \hat{\phi} = 0.045, \qquad \hat{\omega} = 0.000018,$$

$$\hat{\alpha} = 0.080, \qquad \hat{\beta} = 0.900.$$

Dla ostatniej obserwacji w próbie przyjął:

$$r_T = -0.012, \qquad \hat{a}_T^2 = 0.00036, \qquad \hat{\sigma}_T^2 = 0.00025.$$

a) (2 pkt) Na podstawie wyników testów Ljunga–Boxa oceń, czy: (i) w szeregu r_t występuje istotna autokorelacja, (ii) w szeregu r_t^2 występuje zależność świadcząca o zmienności warunkowej. Wyjaśnij, dlaczego model białego szumu o stałej wariancji może być niewystarczający, mimo że test Ljunga–Boxa dla samego szeregu r_t nie prowadzi do odrzucenia hipotezy o braku autokorelacji.

b) (3 pkt) Dla oszacowanego modelu AR(1)–GARCH(1,1):

(i) sprawdź warunek stacjonarności wariancji,

(ii) oblicz wariancję bezwarunkową składnika losowego a_t,

(iii) oblicz prognozę warunkowej wartości oczekiwanej $E(r_{T+1} \mid \mathcal{F}_T)$ oraz prognozę wariancji warunkowej $\hat{\sigma}_{T+1}^2$.

Rozwiązanie

Zakres literatury

Tematyka zadania jest omawiana w szczególności w następujących pozycjach z wykazu literatury:

- A.J. McNeil, R. Frey, P. Embrechts, *Quantitative Risk Management: Concepts, Techniques and Tools*, revised edition, Princeton, 2015: rozdział 3.1 – własności szeregów czasowych stóp zwrotu, w tym grupowanie zmienności; rozdział 4 – szeregi czasowe, w szczególności stacjonarność, autokorelacja, modele ARMA, ARCH/GARCH, prognozowanie, estymacja i ocena dopasowania.
- E.W. Frees, *Regression Modeling with Actuarial and Financial Applications*, Cambridge, 2009: rozdziały 7–9 – szeregi czasowe, w tym stacjonarność, autokorelacja, modele ARMA, ARCH/GARCH, prognozowanie, estymacja i ocena dopasowania.

a) Dla szeregu stóp zwrotu r_t wartość p-value w teście Ljunga–Boxa wynosi:

$$p = 0.211.$$

Na poziomie istotności 0.05 nie ma podstaw do odrzucenia hipotezy o braku autokorelacji. Oznacza to, że w samych stopach zwrotu nie widać istotnej zależności liniowej.

Dla szeregu kwadratów stóp zwrotu r_t^2 otrzymano:

$$p < 0.001.$$

Na poziomie istotności 0.05 odrzucamy hipotezę o braku autokorelacji w kwadratach stóp zwrotu. Wskazuje to na zależność w zmienności, czyli na efekt ARCH/GARCH.

Wniosek:

$$\boxed{\text{Brak istotnej autokorelacji w } r_t \text{ nie oznacza stałej wariancji.}}$$

Szereg może przypominać biały szum w średniej, ale jednocześnie wykazywać grupowanie zmienności. Dlatego model białego szumu o stałej wariancji jest niewystarczający, a zastosowanie modelu GARCH jest uzasadnione.

b) (i) Warunek stacjonarności wariancji.

Dla modelu GARCH(1,1) warunek stacjonarności wariancji ma postać:

$$\alpha + \beta < 1.$$

W tym przypadku:

$$\hat{\alpha} + \hat{\beta} = 0.080 + 0.900 = 0.980 < 1.$$

Zatem:

$$\boxed{\text{warunek stacjonarności wariancji jest spełniony.}}$$

Jednocześnie suma parametrów jest bliska jedności, co oznacza dużą trwałość szoków zmienności.

(ii) Wariancja bezwarunkowa.

Wariancja bezwarunkowa składnika losowego a_t wynosi:

$$\mathrm{Var}(a_t) = \frac{\omega}{1-\alpha-\beta}.$$

Po podstawieniu oszacowań:

$$\widehat{\mathrm{Var}}(a_t) = \frac{0.000018}{1-0.080-0.900} = 0.0009.$$

Zatem:

$$\boxed{\widehat{\mathrm{Var}}(a_t) = 0.0009.}$$

Odpowiadające bezwarunkowe odchylenie standardowe wynosi 0.03, czyli około 3% dziennie.

(iii) Prognozy na okres $T+1$.

Prognoza warunkowej wartości oczekiwanej wynosi:

$$E(r_{T+1} \mid \mathcal{F}_T) = \hat{\mu} + \hat{\phi} r_T.$$

Po podstawieniu:

$$E(r_{T+1} \mid \mathcal{F}_T) = 0.00012 + 0.045 \cdot (-0.012) = -0.00042.$$

Zatem:

$$\boxed{E(r_{T+1} \mid \mathcal{F}_T) = -0.00042.}$$

Prognoza wariancji warunkowej wynosi:

$$\hat{\sigma}^2_{T+1} = \hat{\omega} + \hat{\alpha}\hat{a}^2_T + \hat{\beta}\hat{\sigma}^2_T.$$

Po podstawieniu:

$$\hat{\sigma}^2_{T+1} = 0.000018 + 0.080 \cdot 0.00036 + 0.900 \cdot 0.00025 = 0.0002718.$$

Zatem:

$$\boxed{\hat{\sigma}^2_{T+1} = 0.0002718.}$$

Odpowiadające odchylenie standardowe warunkowe wynosi:

$$\hat{\sigma}_{T+1} \approx \sqrt{0.0002718} \approx 0.0165,$$

czyli około 1.65% dziennie.

Zadanie 6

W pewnym portfelu ubezpieczeń komunikacyjnych porównano dwa modele taryfikacyjne: model A oraz model B. Dla każdej polisy znany jest rzeczywisty koszt ryzyka w okresie testowym Y_i oraz składki czyste

$$\hat{\mu}_i^A, \qquad \hat{\mu}_i^B,$$

prognozowane odpowiednio przez modele A i B.

Dane przedstawione są w poniższej tabeli. Wszystkie wartości podano w tych samych jednostkach pieniężnych.

Polisa	Y_i	$\hat{\mu}_i^A$	$\hat{\mu}_i^B$
1	80	90	100
2	130	120	240
3	160	180	180
4	180	210	140
5	230	260	300
6	300	290	260
7	400	350	430
8	520	500	350

Dla obu modeli zachodzi globalny warunek bilansu:

$$\sum_i Y_i = \sum_i \hat{\mu}_i^A = \sum_i \hat{\mu}_i^B = 2000.$$

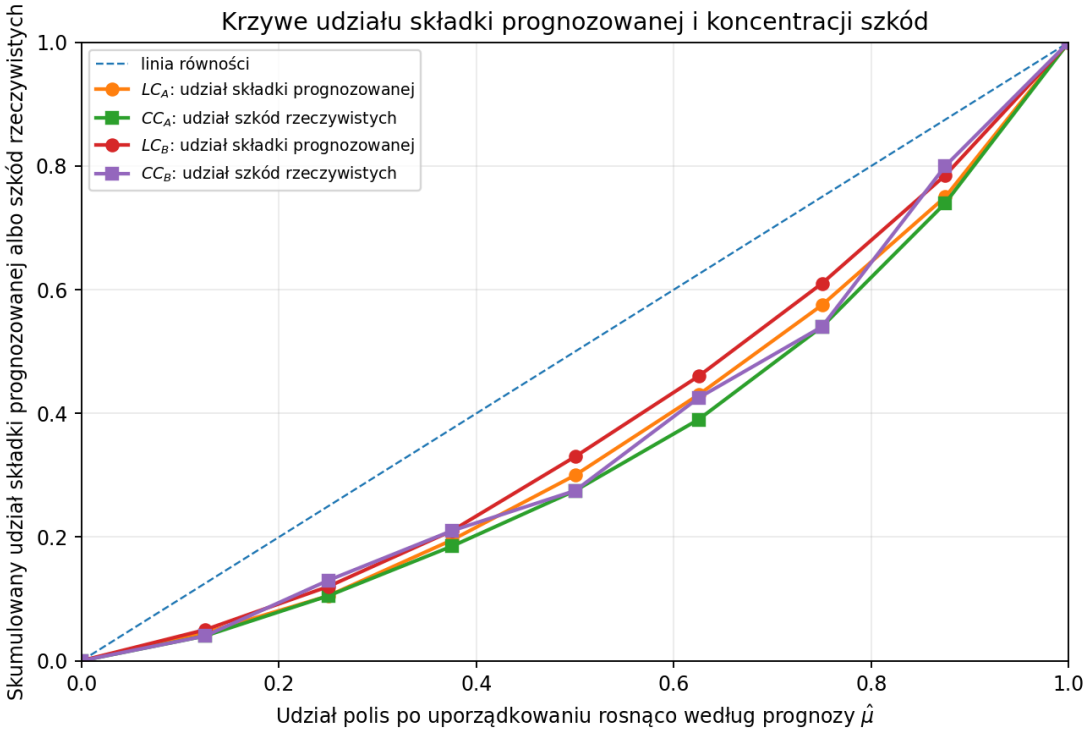
W tym zadaniu przyjmij następujący sposób konstrukcji krzywych. Dla wybranego modelu polisy porządkujemy rosnąco według prognozowanej składki czystej $\hat{\mu}_i$.

Dla $\alpha_k = k/n$, gdzie $k = 0, 1, \ldots, n$, definiujemy dwie wielkości:

$$LC(\alpha_k) = \frac{\sum_{j=1}^{k} \hat{\mu}_{(j)}}{\sum_{i=1}^{n} \hat{\mu}_i}, \qquad CC(\alpha_k) = \frac{\sum_{j=1}^{k} Y_{(j)}}{\sum_{i=1}^{n} Y_i}.$$

W powyższych wzorach indeks (j) oznacza j-tą polisę po uporządkowaniu rosnąco według $\hat{\mu}_i$.

Na rysunku 6.1 przedstawiono krzywe LC oraz CC dla obu modeli.



Rysunek 6.1: Krzywe udziału składki prognozowanej LC i koncentracji szkód rzeczywistych CC dla modeli taryfikacyjnych A oraz B.

Do obliczenia współczynnika Giniego dla krzywej LC przyjmij:

$$A_{LC}^{A} = 0.3625, \qquad A_{LC}^{B} = 0.3831,$$

gdzie A_{LC}^{A} i A_{LC}^{B} oznacza pole pod krzywą LC odpowiednio dla modelu A i B.

a) (1 pkt) Dla każdego z modeli wyznacz wartości $LC(0.5)$ oraz $CC(0.5)$, czyli wartości krzywych dla 50% polis o najniższych prognozowanych składkach czystych.

b) (2 pkt) Oblicz współczynniki Giniego dla modeli A oraz B. Który model silniej różnicuje składki pomiędzy ryzykami? Czy sam wyższy współczynnik Giniego wystarcza do stwierdzenia, że model taryfikacyjny jest lepszy? Uzasadnij odpowiedź.

c) (2 pkt) Na podstawie rysunku 6.1 porównaj modele A i B z punktu widzenia zgodności krzywej CC z krzywą LC. Wyjaśnij, co oznacza sytuacja, w której dla pewnego zakresu α krzywa CC leży wyraźnie poniżej albo powyżej krzywej LC.

Rozwiązanie

Zakres literatury

Tematyka zadania jest omawiana w szczególności w następujących pozycjach z wykazu literatury:

- M. Denuit, D. Hainaut, J. Trufin, *Effective Statistical Learning Methods for Actuaries II*, Springer, 2021: rozdział 6.3, w szczególności podrozdziały 6.3.4–6.3.5 dotyczące krzywych koncentracji, krzywych Lorenza oraz oceny jakości predyktora taryfowego; podrozdział 6.3.6 dotyczący porównywania predyktorów.
- E.W. Frees, G. Meyers, A.D. Cummings, prace dotyczące wykorzystania indeksu Giniego w taryfikacji ubezpieczeniowej, przywoływane w bibliografii rozdziału 6 książki Denuit, Hainaut, Trufin.

a) Dla modelu A kolejność polis według rosnących wartości $\hat{\mu}_i^A$ jest taka sama jak w tabeli. Dla 50% polis o najniższych prognozowanych składkach bierzemy polisy $1, 2, 3, 4$. Otrzymujemy:

$$LC_A(0.5) = \frac{90 + 120 + 180 + 210}{2000} = 0.300,$$

$$CC_A(0.5) = \frac{80 + 130 + 160 + 180}{2000} = 0.275.$$

Dla modelu B cztery najniższe prognozy odpowiadają polisom:

$$1,\ 4,\ 3,\ 2.$$

Stąd:

$$LC_B(0.5) = \frac{100 + 140 + 180 + 240}{2000} = 0.330,$$

$$CC_B(0.5) = \frac{80 + 180 + 160 + 130}{2000} = 0.275.$$

Zatem:

$$\boxed{LC_A(0.5) = 0.300, \quad CC_A(0.5) = 0.275,}$$

$$\boxed{LC_B(0.5) = 0.330, \quad CC_B(0.5) = 0.275.}$$

W obu modelach w dolnej połowie portfela udział szkód rzeczywistych jest niższy niż udział składki prognozowanej.

b) Współczynnik Giniego dla krzywej LC liczony jest jako:

$$G = 2(0.5 - A_{LC}).$$

Dla modelu A:

$$G_A = 2(0.5 - 0.3625) = 0.275.$$

Dla modelu B:

$$G_B = 2(0.5 - 0.3831) = 0.2338.$$

Zatem:

$$\boxed{G_A = 0.275, \qquad G_B = 0.2338.}$$

Model A silniej różnicuje prognozowane składki między ryzykami, ponieważ ma wyższy współczynnik Giniego.

Sam wyższy współczynnik Giniego nie wystarcza jednak do stwierdzenia, że model taryfikacyjny jest lepszy. Współczynnik Giniego dla krzywej LC mierzy przede wszystkim zróżnicowanie prognozowanych składek. Nie przesądza natomiast, czy to zróżnicowanie jest zgodne z rzeczywistą strukturą szkód. Dlatego należy oceniać również zgodność krzywej koncentracji szkód CC z krzywą udziału składki prognozowanej LC, a także kalibrację i stabilność modelu.

c) Na podstawie rysunku widać, że dla obu modeli krzywe LC i CC są stosunkowo blisko siebie, ale nie pokrywają się idealnie.

Dla modelu A krzywa CC_A przez większą część zakresu leży nieco poniżej krzywej LC_A. Oznacza to, że wśród polis o niższych prognozowanych składkach udział rzeczywistych szkód jest nieco niższy niż udział przypisanej im składki prognozowanej.

Dla modelu B rozbieżności między LC_B i CC_B są widoczne szczególnie w środkowej części portfela. W tej części portfela udział rzeczywistych szkód jest niższy niż udział składki prognozowanej, co może wskazywać na relatywne przeszacowanie ryzyka wśród polis o niższych i średnich prognozach.

Ogólnie:

$$CC(\alpha) < LC(\alpha)$$

oznacza, że w grupie $100\alpha\%$ polis o najniższych prognozowanych składkach udział szkód rzeczywistych jest niższy niż udział składki prognozowanej. Taka część portfela jest względnie przeszacowana.

Natomiast:

$$CC(\alpha) > LC(\alpha)$$

oznacza, że udział szkód rzeczywistych jest wyższy niż udział składki prognozowanej. Taka część portfela jest względnie niedoszacowana.

Wniosek:

dobry model taryfikacyjny powinien nie tylko różnicować składki, ale robić to zgodnie z rzeczywistym kosztem ryzyka.

Zadanie 7

Aktuariusz przygotowuje model rezerw szkodowych dla portfela ubezpieczeń komunikacyjnych. Model ma zostać wykorzystany przy zamknięciu roku oraz w materiale dla zarządu.

W trakcie prac ustalono, że dane za lata 2019–2020 są częściowo niekompletne z powodu migracji systemu, część informacji o datach zgłoszenia pochodzi od zewnętrznego dostawcy, a model zmieniono względem poprzedniego roku: zmodyfikowano segmentację portfela oraz sposób traktowania wysokich szkód. Dla nowej wersji modelu wykonano testy wrażliwości na inflację szkód, liczbę wysokich szkód oraz wybór okresu kalibracji.

Rozważ tę sytuację z punktu widzenia Krajowego Standardu Aktuarialnego Polskiego Stowarzyszenia Aktuariuszy *Praktyka aktuarialna*.

a) (3 pkt) Wskaż, jakie działania powinny zostać wykonane przed użyciem zmodyfikowanego modelu do wyznaczenia rezerw.

b) (2 pkt) Wskaż, jakie informacje powinny znaleźć się w raporcie dla zarządu, aby użytkownicy właściwie rozumieli wyniki.

Odpowiedzi

Zakres literatury

Tematyka zadania jest omawiana w szczególności w następujących pozycjach z wykazu literatury:

- *Krajowy Standard Aktuarialny Polskiego Stowarzyszenia Aktuariuszy – Praktyka Aktuarialna*, 2022: w szczególności podrozdziały 2.3, 2.5, 2.7, 2.10 oraz 3.1–3.2, dotyczące polegania na informacjach od stron trzecich, jakości i braków danych, założeń i metodologii, zarządzania modelami, walidacji, testów wrażliwości oraz raportowania.
- *Actuarial Aspects of ERM for Insurance Companies*, 2016: podrozdziały 3.5.2–3.5.3, dotyczące kalibracji i walidacji modeli aktuarialnych, analizy wrażliwości oraz danych wykorzystywanych w modelach.

a) Przed wykorzystaniem zmodyfikowanego modelu do wyznaczenia rezerw aktuariusz powinien przede wszystkim upewnić się, że model, dane i proces obliczeniowy są odpowiednie do celu pracy, czyli do wyznaczenia rezerw na potrzeby zamknięcia roku.

W szczególności powinien:

- ocenić jakość danych, zwłaszcza kompletność i spójność danych za lata 2019–2020;
- zbadać wpływ migracji systemu na dane szkodowe, np. przez porównania historyczne, testy racjonalności i uzgodnienia z innymi źródłami;
- ocenić wiarygodność danych o datach zgłoszenia pochodzących od zewnętrznego dostawcy oraz określić, w jakim zakresie na tych danych polega;
- określić sposób postępowania z brakami danych, np. korektę, uzupełnienie, ograniczenie okresu kalibracji, dodatkowy margines ostrożności albo analizę wrażliwości;
- zwalidować zmodyfikowany model, sprawdzając, czy jest odpowiedni do celu, spełnia specyfikację i daje powtarzalne oraz racjonalne wyniki;
- porównać nową wersję modelu z wersją z poprzedniego roku i wyjaśnić wpływ zmian w segmentacji portfela oraz sposobie traktowania wysokich szkód;
- zidentyfikować ryzyko modelu, w tym ryzyko błędnej segmentacji, nieadekwatnego traktowania wysokich szkód, niewłaściwego okresu kalibracji i nieadekwatnych założeń inflacyjnych;
- zapewnić kontrolę zmian i kontrolę uruchomień modelu, tak aby wiadomo było, jaka wersja modelu, jakie dane i jakie założenia zostały użyte;
- wykorzystać wykonane testy wrażliwości do oceny stabilności wyniku rezerw.

Wniosek:

model może być użyty do wyznaczenia rezerw dopiero po ocenie danych, walidacji modelu, udokumentowaniu zmian i kontroli procesu obliczeniowego.

b) Raport dla zarządu powinien być zwięzły, ale powinien umożliwiać właściwe zrozumienie wyniku, jego ograniczeń i znaczenia decyzyjnego.

W raporcie powinny znaleźć się w szczególności:

- cel raportu i zamierzone wykorzystanie wyników, czyli wyznaczenie rezerw na potrzeby zamknięcia roku;
- podstawowe wyniki modelu oraz ich potencjalny wpływ na poziom rezerw;
- opis wykorzystanych danych, w tym informacja o niekompletności danych za lata 2019–2020;
- informacja o danych pochodzących od zewnętrznego dostawcy i zakresie polegania na tych danych;
- opis najważniejszych założeń, w szczególności dotyczących inflacji szkód, wysokich szkód oraz okresu kalibracji;
- opis zmian względem poprzedniego roku, zwłaszcza zmian segmentacji portfela i traktowania wysokich szkód;
- informacja o walidacji modelu i kontrolach wykonanych przed użyciem modelu;
- wyniki testów wrażliwości i wskazanie, które założenia mają największy wpływ na rezerwy;
- ograniczenia danych i modelu oraz niepewność wyników, wraz z ich możliwymi konsekwencjami dla decyzji zarządu;
- informacja o zastosowanych korektach eksperckich, marginesach ostrożności lub odstępstwach od standardowej procedury, jeżeli wystąpiły.

Wniosek:

zarząd powinien otrzymać nie tylko wartość rezerw, lecz także informację o danych, założeniach, niepewności, ograniczeniach i wynikach testów wrażliwości.

Zadanie 8

W pewnym portfelu ubezpieczeń komunikacyjnych zbudowano model predykcyjny rocznej częstości szkód:

$$\hat{f}(x) = \widehat{E}[N \mid x],$$

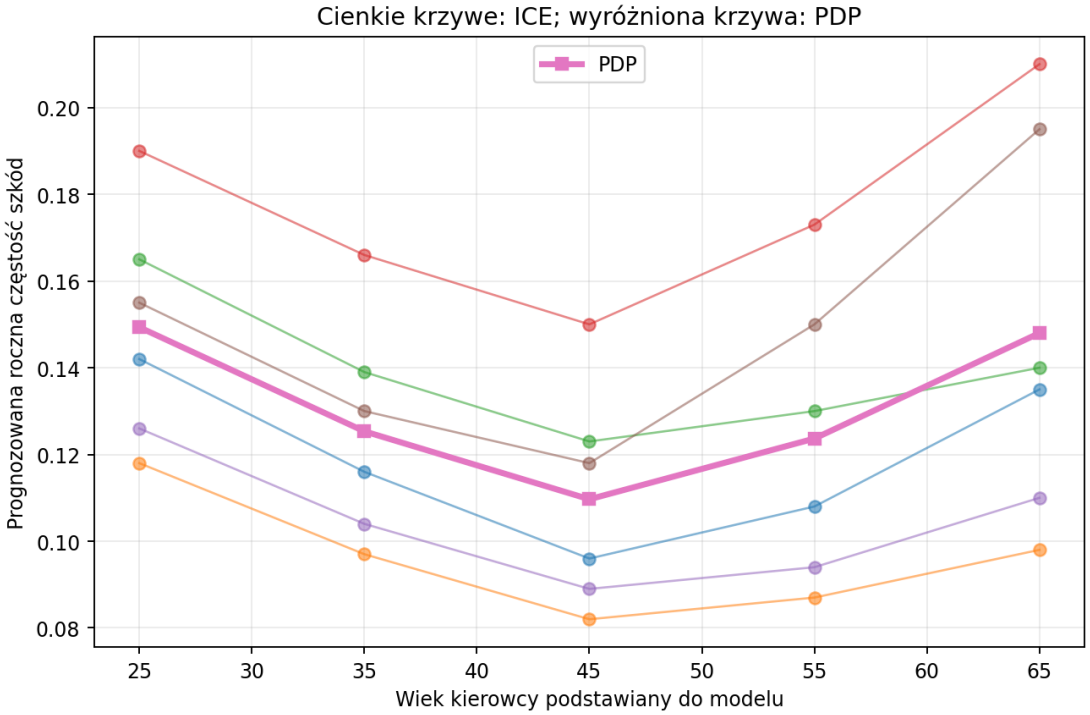
gdzie N oznacza roczną liczbę szkód. Aktuariusz analizuje wpływ zmiennej *DriverAge* na prognozowaną częstość szkód za pomocą krzywych ICE (*individual conditional expectation*) oraz PDP (*partial dependence plot*).

Dla sześciu polis z próby testowej otrzymano następujące wartości ICE dla siatki wieku

$$a = 25,\ 35,\ 45,\ 55,\ 65:$$

Polisa	$a = 25$	$a = 35$	$a = 45$	$a = 55$	$a = 65$
1	0.142	0.116	0.096	0.108	0.135
2	0.118	0.097	0.082	0.087	0.098
3	0.165	0.139	0.123	0.130	0.140
4	0.190	0.166	0.150	0.173	0.210
5	0.126	0.104	0.089	0.094	0.110
6	0.155	0.130	0.118	0.150	0.195

Na rysunku 8.1 przedstawiono odpowiadające tym wartościom krzywe ICE oraz krzywą PDP.



Rysunek 8.1: Krzywe ICE oraz PDP dla zmiennej *DriverAge*.

Dodatkowo wiadomo, że w analizowanym portfelu *DriverAge* jest silnie dodatnio skorelowany ze stażem klienta w zakładzie ubezpieczeń. Młodzi kierowcy bardzo rzadko mają staż przekraczający 10 lat, natomiast w starszych grupach wiekowych długi staż występuje często.

a) (1 pkt) Na podstawie wartości ICE z tabeli oblicz wartości krzywej PDP dla wieku:

$$a = 25, \qquad a = 45, \qquad a = 65.$$

b) (2 pkt) Porównaj informację dostarczaną przez krzywą PDP z informacją dostarczaną przez krzywe ICE. Wskaż, czy wpływ wieku kierowcy jest taki sam dla wszystkich polis, czy też widoczna jest heterogeniczność efektu. Wyjaśnij, o czym może świadczyć rozbieżny kształt poszczególnych krzywych ICE.

c) (2 pkt) Wyjaśnij, dlaczego interpretacja PDP może być myląca, gdy analizowana zmienna jest silnie skorelowana z innymi cechami. Odnieś odpowiedź do podanej informacji o zależności między *DriverAge* a stażem klienta. Zaproponuj jedną metodę uzupełniającą analizę PDP w takiej sytuacji.

Rozwiązanie

Zakres literatury

Tematyka zadania jest omawiana w szczególności w następujących pozycjach z wykazu literatury:

- M. Denuit, D. Hainaut, J. Trufin, *Effective Statistical Learning Methods for Actuaries II*, Springer, 2020: rozdziały 3–5, w szczególności modele drzewiaste, metody zespołowe, istotność predykcyjna zmiennych, *partial dependence plots* oraz statystyka H Friedmana.
- G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, rozdział 8 – drzewa regresyjne i klasyfikacyjne, bagging, random forest oraz metody interpretacji modeli oparte na analizie wpływu zmiennych.

a) Krzywa PDP dla danej wartości wieku jest średnią z wartości ICE dla poszczególnych polis.

Dla $a = 25$:

$$PD(25) = \frac{0.142 + 0.118 + 0.165 + 0.190 + 0.126 + 0.155}{6} \approx 0.149.$$

Dla $a = 45$:

$$PD(45) = \frac{0.096 + 0.082 + 0.123 + 0.150 + 0.089 + 0.118}{6} \approx 0.110.$$

Dla $a = 65$:

$$PD(65) = \frac{0.135 + 0.098 + 0.140 + 0.210 + 0.110 + 0.195}{6} \approx 0.148.$$

Zatem:

$$\boxed{PD(25) \approx 0.149, \qquad PD(45) \approx 0.110, \qquad PD(65) \approx 0.148.}$$

b) Krzywa PDP pokazuje przeciętny wpływ wieku kierowcy na prognozowaną częstość szkód, po uśrednieniu prognoz po analizowanych polisach. Na podstawie wartości z punktu a) oraz rysunku widać zależność zbliżoną do kształtu litery U: prognozowana częstość szkód jest wyższa dla młodszych kierowców, maleje w okolicach wieku 45 lat, a następnie ponownie rośnie dla starszych kierowców.

Krzywe ICE pokazują natomiast zmianę prognozy dla pojedynczych polis. Dzięki temu ujawniają, czy efekt wieku jest podobny dla wszystkich obserwacji. W tym przypadku widać wspólny ogólny wzorzec, ale także różnice poziomu i nachylenia poszczególnych krzywych.

Oznacza to, że wpływ wieku nie jest identyczny dla wszystkich polis:

$$\boxed{\text{widoczna jest heterogeniczność efektu wieku kierowcy.}}$$

Rozbieżny kształt krzywych ICE może świadczyć o interakcjach między wiekiem kierowcy a innymi cechami polisy, np. stażem klienta, typem pojazdu, regionem, historią szkodową albo innymi zmiennymi użytymi w modelu. Innymi słowy, wpływ wieku może zależeć od profilu konkretnej polisy.

c) Interpretacja PDP może być myląca, gdy analizowana zmienna jest silnie skorelowana z innymi cechami, ponieważ przy konstrukcji PDP rozważa się również sztuczne kombinacje zmiennych,

które mogą być rzadkie albo nierealistyczne w rzeczywistym portfelu.

W tym zadaniu *DriverAge* jest silnie dodatnio skorelowany ze stażem klienta. Młodzi kierowcy rzadko mają staż przekraczający 10 lat, a starsi kierowcy często mają długi staż. Przy analizie PDP model może jednak oceniać prognozy dla kombinacji takich jak młody kierowca z bardzo długim stażem albo starszy kierowca z bardzo krótkim stażem. Takie kombinacje mogą być słabo reprezentowane w danych.

W takiej sytuacji krzywa PDP może częściowo odzwierciedlać ekstrapolację modelu poza typowy zakres danych, a nie wyłącznie rzeczywisty efekt wieku obserwowany w portfelu.

Możliwa metoda uzupełniająca:

analiza ICE lub warunkowe PDP liczone w podgrupach stażu klienta.

Można również rozważyć metodę ALE (*accumulated local effects*), która jest mniej wrażliwa na korelację między zmiennymi, ponieważ opiera się bardziej na lokalnych zmianach w obszarach faktycznie obserwowanych w danych.

Wniosek:

PDP należy interpretować ostrożnie, gdy zmienne objaśniające są silnie skorelowane.

Zadanie 9

W portfelu ubezpieczeń na życie obserwowano grupę 50 ubezpieczonych od wieku 60 lat. Niech T oznacza czas, w latach, od wieku 60 lat do zgonu. Część obserwacji zakończyła się przed wystąpieniem zgonu, np. z powodu końca okresu obserwacji lub utraty obserwacji; takie obserwacje traktowano jako prawostronnie cenzurowane.

Dla kolejnych momentów zgonu otrzymano następujące dane:

Czas zgonu $t_{(j)}$	Liczba zgonów d_j	Liczba cenzurowań po $t_{(j)}$
2	3	2
4	4	3
6	5	3
8	4	2
10	3	5

W kolumnie "Liczba cenzurowań po $t_{(j)}$" podano liczbę obserwacji cenzurowanych po danym czasie zgonu, ale przed kolejnym czasem zgonu. Przyjmij, że przed czasem $t = 2$ nie wystąpiły cenzurowania.

a) (2 pkt) Oblicz estymator Kaplana–Meiera funkcji przeżycia dla:

$$t = 8, \qquad t = 10.$$

Zinterpretuj wartość $\widehat{S}(8)$.

b) (2 pkt) Dla punktu $t = 8$ oblicz wariancję estymatora Kaplana–Meiera, stosując formułę Greenwooda. Następnie wyznacz przybliżony 95% przedział ufności dla $S(8)$, korzystając z normalnego przybliżenia.

c) (1 pkt) Wyjaśnij, jakie założenie dotyczące obserwacji cenzurowanych jest istotne przy stosowaniu estymatora Kaplana–Meiera. Co mogłoby się stać z oszacowaniem funkcji przeżycia, gdyby cenzurowanie było związane z pogorszonym stanem zdrowia ubezpieczonych?

Rozwiązanie

Zakres literatury

Tematyka zadania jest omawiana w szczególności w następujących pozycjach z wykazu literatury:

- D.C.M. Dickson, M.R. Hardy, H.R. Waters, *Actuarial Mathematics for Life Contingent Risks*, Cambridge, 2020: rozdział 18, w szczególności podrozdziały 18.3.2 oraz 18.4.3, dotyczące prawostronnego cenzurowania, estymatora Kaplana–Meiera oraz formuły Greenwooda.
- S.A. Klugman, H.H. Panjer, G.E. Willmot, *Loss Models: From Data to Decisions*, Wiley, 2019: rozdziały 14.1–14.5, dotyczące empirycznej estymacji funkcji przeżycia dla danych cenzurowanych oraz zastosowania estymatorów Kaplana–Meiera i Nelsona–Aalena.

a) Najpierw należy odtworzyć liczby w ryzyku r_j tuż przed kolejnymi momentami zgonu. Ponieważ przed czasem $t = 2$ nie było cenzurowań, mamy:

$$r_1 = 50.$$

Po każdym czasie zgonu odejmujemy zarówno zgony, jak i cenzurowania występujące przed następnym czasem zgonu. Otrzymujemy:

$t_{(j)}$	2	4	6	8	10
r_j	50	45	38	30	24
d_j	3	4	5	4	3

Estymator Kaplana–Meiera dla $t = 8$ wynosi:

$$\widehat{S}(8) = \left(1 - \frac{3}{50}\right)\left(1 - \frac{4}{45}\right)\left(1 - \frac{5}{38}\right)\left(1 - \frac{4}{30}\right) \approx 0.645.$$

Dla $t = 10$:

$$\widehat{S}(10) = \widehat{S}(8)\left(1 - \frac{3}{24}\right) \approx 0.564.$$

Zatem:

$$\boxed{\widehat{S}(8) \approx 0.645, \qquad \widehat{S}(10) \approx 0.564.}$$

Interpretacja: $\widehat{S}(8) \approx 0.645$ oznacza, że szacowane prawdopodobieństwo przeżycia od wieku 60 lat do wieku 68 lat wynosi około 0.645.

b) Dla $t = 8$, zgodnie z formułą Greenwooda:

$$\widehat{\mathrm{Var}}\{\widehat{S}(8)\} = \widehat{S}(8)^2 \sum_{j:t_{(j)} \le 8} \frac{d_j}{r_j(r_j - d_j)}.$$

Po podstawieniu:

$$\sum_{j:t_{(j)} \le 8} \frac{d_j}{r_j(r_j - d_j)} = \frac{3}{50 \cdot 47} + \frac{4}{45 \cdot 41} + \frac{5}{38 \cdot 33} + \frac{4}{30 \cdot 26} \approx 0.01256.$$

Ponieważ $\widehat{S}(8) \approx 0.6446$, otrzymujemy:

$$\widehat{\mathrm{Var}}\{\widehat{S}(8)\} \approx 0.6446^2 \cdot 0.01256 \approx 0.00522.$$

Zatem:

$$\boxed{\widehat{\mathrm{Var}}\{\widehat{S}(8)\} \approx 0.00522.}$$

Błąd standardowy:

$$\widehat{SE}\{\widehat{S}(8)\} = \sqrt{0.00522} \approx 0.0723.$$

Przybliżony 95% przedział ufności, korzystając z normalnego przybliżenia, wynosi:

$$\widehat{S}(8) \pm 1.96\, \widehat{SE}\{\widehat{S}(8)\}.$$

Stąd:

$$0.6446 \pm 1.96 \cdot 0.0723 \approx 0.6446 \pm 0.1416.$$

Ostatecznie:

$$\boxed{CI_{0.95}(S(8)) \approx [0.503,\ 0.786].}$$

c) Przy stosowaniu estymatora Kaplana–Meiera istotne jest założenie nieinformacyjnego cenzurowania. Oznacza ono, że sam fakt cenzurowania nie powinien nieść informacji o dalszym ryzyku zgonu osoby cenzurowanej, po uwzględnieniu dotychczasowego przeżycia.

Jeżeli cenzurowanie byłoby związane z pogorszonym stanem zdrowia, osoby o podwyższonym ryzyku zgonu byłyby częściej usuwane z obserwacji przed zaobserwowaniem zgonu. W takiej sytuacji estymator Kaplana–Meiera mógłby zawyżać funkcję przeżycia, czyli prowadzić do zbyt optymistycznej oceny prawdopodobieństwa przeżycia.

$$\boxed{\text{Informacyjne cenzurowanie może prowadzić do obciążonej oceny funkcji przeżycia.}}$$

Zadanie 10

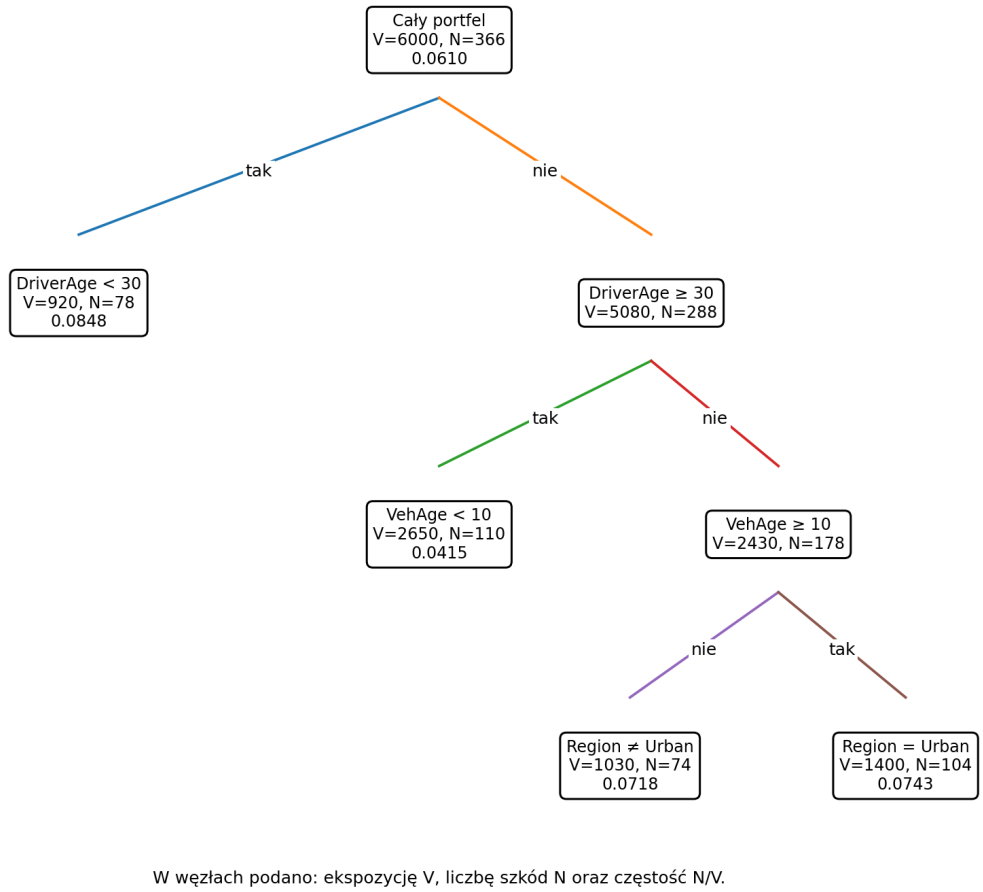
W pewnym portfelu ubezpieczeń komunikacyjnych modelowano roczną liczbę szkód N_i z uwzględnieniem ekspozycji v_i, wyrażonej w latach. Rozważono poissonowskie drzewo regresyjne. Do budowy drzewa wykorzystano zmienne:

$$DriverAge \quad \text{– wiek kierowcy,}$$

$$VehAge \quad \text{– wiek pojazdu,}$$

$$Region \quad \text{– region użytkowania pojazdu.}$$

Na rysunku 10.1 przedstawiono poddrzewo, które należy wykorzystać do obliczenia prognoz w punkcie a).



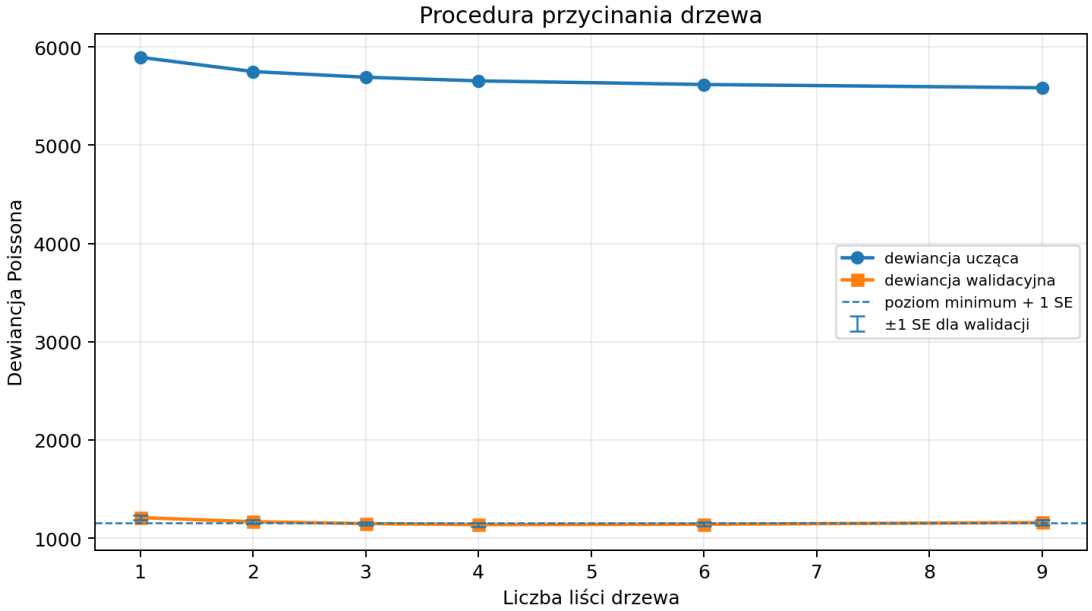
Rysunek 10.1: Poddrzewo używane do wyznaczenia prognoz w punkcie a).

Dla liści drzewa z rysunku 10.1 otrzymano następujące wartości:

Liść	Warunek	Ekspozycja $\sum v_i$	Liczba szkód $\sum N_i$
L_1	$DriverAge < 30$	920	78
L_2	$DriverAge \geq 30,\ VehAge < 10$	2650	110
L_3	$DriverAge \geq 30,\ VehAge \geq 10,\ Region \neq Urban$	1030	74
L_4	$DriverAge \geq 30,\ VehAge \geq 10,\ Region = Urban$	1400	104

Następnie przeprowadzono procedurę przycinania większego drzewa. Rozważano poddrzewa o różnej liczbie liści. W tabeli oraz na rysunku 10.2 przedstawiono dewiancję uczącą i walidacyjną.

Liczba liści	Dewiancja ucząca	Dewiancja walidacyjna	SE walidacji
1	5895	1210	22
2	5750	1168	18
3	5692	1149	17
4	5655	1138	18
6	5618	1142	20
9	5585	1160	25



Rysunek 10.2: Dewiancja ucząca i walidacyjna dla poddrzew o różnej liczbie liści.

a) (2 pkt) Na podstawie drzewa z rysunku 10.1 wyznacz prognozowaną liczbę szkód dla polisy:

(i) o ekspozycji $v = 0.5$, jeżeli kierowca ma 42 lata, pojazd ma 12 lat, a region użytkowania pojazdu to *Urban*,

(ii) o ekspozycji $v = 0.5$, jeżeli kierowca ma 22 lata, pojazd ma 12 lat, a region użytkowania pojazdu to *Urban*.

b) (1 pkt) Na podstawie tabeli wyników walidacyjnych i rysunku 10.2 wybierz liczbę liści drzewa, którą przyjąłbyś do dalszego zastosowania. Zastosuj regułę jednego błędu standardowego.

c) (2 pkt) Wyjaśnij, jak należy interpretować pierwszy podział drzewa $DriverAge < 30$. Podaj jedną zaletę i jedno ograniczenie drzewa regresyjnego w porównaniu z klasycznym modelem GLM w zadaniu taryfikacyjnym.

Rozwiązanie

Zakres literatury

Tematyka zadania jest omawiana w szczególności w następujących pozycjach z wykazu literatury:

- M. Denuit, D. Hainaut, J. Trufin, *Effective Statistical Learning Methods for Actuaries II*, Springer, 2020: rozdział 2 – moc predykcyjna modelu przy funkcji straty Poissona; rozdziały 3–5 – drzewa regresyjne, algorytm budowy drzewa, pruning, bagging, boosting, random forest oraz istotność predykcyjna zmiennych.
- M.V. Wüthrich, C. Buser, *Data Analytics for Non-Life Insurance Pricing*, 2020: rozdział 6 – poissonowskie drzewa regresyjne i drzewa klasyfikacyjne; rozdziały 7.2–7.4 – bagging, boosting i random forest dla poissonowskich drzew regresyjnych.
- G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, 2021: rozdział 8 – drzewa regresyjne i klasyfikacyjne, pruning, bagging, boosting i random forest.

a) W liściu drzewa prognozowana częstość szkód jest równa:

$$\hat{\lambda}_L = \frac{\sum_{i \in L} N_i}{\sum_{i \in L} v_i}.$$

Dla polisy o ekspozycji v prognozowana liczba szkód wynosi:

$$\widehat{E[N]} = v\hat{\lambda}_L.$$

(i) Kierowca ma 42 lata, pojazd ma 12 lat, a region to $Urban$, więc polisa trafia do liścia:

$$L_4: \quad DriverAge \geq 30, \quad VehAge \geq 10, \quad Region = Urban.$$

Dla tego liścia:

$$\hat{\lambda}_{L_4} = \frac{104}{1400} \approx 0.0743.$$

Przy ekspozycji $v = 0.5$:

$$\widehat{E[N]} = 0.5 \cdot 0.0743 \approx 0.0371.$$

Zatem:

$$\boxed{\widehat{E[N]} \approx 0.037.}$$

(ii) Kierowca ma 22 lata, więc polisa trafia do liścia:

$$L_1: \quad DriverAge < 30.$$

Dla tego liścia:

$$\hat{\lambda}_{L_1} = \frac{78}{920} \approx 0.0848.$$

Przy ekspozycji $v = 0.5$:

$$\widehat{E[N]} = 0.5 \cdot 0.0848 \approx 0.0424.$$

Zatem:

$$\boxed{\widehat{E[N]} \approx 0.042.}$$

b) Najmniejsza dewiancja walidacyjna występuje dla drzewa z 4 liśćmi:

$$D_{\min} = 1138, \qquad SE = 18.$$

Reguła jednego błędu standardowego wybiera najprostsze drzewo, którego dewiancja walidacyjna nie przekracza:

$$D_{\min} + SE = 1138 + 18 = 1156.$$

Drzewo z 3 liśćmi spełnia ten warunek:

$$1149 \leq 1156,$$

natomiast drzewo z 2 liśćmi już go nie spełnia:

$$1168 > 1156.$$

Zatem zgodnie z regułą jednego błędu standardowego należy wybrać:

$$\boxed{\text{drzewo z 3 liśćmi.}}$$

c) Pierwszy podział:

$$DriverAge < 30$$

oznacza, że algorytm uznał wiek kierowcy za najważniejszą zmienną na pierwszym etapie segmentacji portfela. Kierowcy młodsi niż 30 lat zostali wydzieleni jako osobna grupa ryzyka.

Dla tej grupy częstość szkód wynosi:

$$\frac{78}{920} \approx 0.0848.$$

Dla pozostałych kierowców, przed dalszymi podziałami, częstość wynosi:

$$\frac{288}{5080} \approx 0.0567.$$

Pierwszy podział można więc interpretować jako wskazanie, że młodsi kierowcy mają wyższą częstość szkód niż pozostali kierowcy w analizowanym portfelu.

Zaleta drzewa regresyjnego:

$$\boxed{\text{drzewo automatycznie wykrywa progi, nieliniowości i interakcje między zmiennymi.}}$$

W tym przykładzie efekt regionu pojawia się dopiero dla kierowców w wieku co najmniej 30 lat i pojazdów co najmniej 10-letnich.

Ograniczenie drzewa regresyjnego:

$$\boxed{\text{predykcje są skokowe, a struktura drzewa może być niestabilna.}}$$

Niewielka zmiana danych może prowadzić do innego podziału, a prognoza jest stała w całym liściu. W porównaniu z klasycznym GLM drzewo jest łatwe do interpretacji segmentacyjnej, ale może gorzej opisywać gładkie zależności i wymaga kontroli złożoności, np. przez przycinanie.

Sesja egzaminacyjna w dniu 26 maja 2026 r.

Modelowanie

Arkusz ocen

Zadanie nr	Punktacja
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	