Komisja Egzaminacyjna dla Aktuariuszy

LXXXII Egzamin dla Aktuariuszy

Sesja egzaminacyjna w dniu 3 marca 2020 r.

Modelowanie

Imię i nazwisko osoby egzaminowanej: .....

Czas trwania egzaminu: 120 minut

Uwagi

a) W prezentowanych wynikach separatorem dziesiętnym (znakiem dziesiętnym) jest kropka „ .”.

b) W prezentowanych wynikach oszacowań uogólnionych modeli liniowych (GLM):
- 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,
- Residual standard error – odchylenie standardowe reszt .

c) W zadaniach wartość zagrożona na poziomie ufności α jest definiowana jako kwantyl rzędu α rozkładu odpowiedniej zmiennej losowej, tzn.

$$VaR_\alpha(X) = \inf\{x: F_X(x) \geq \alpha\} .$$

d) W zadaniach zastosowano następujące oznaczenia:
$E(X)$ – wartość oczekiwana
$D(X)$ – odchylenie standardowe

e) Wartości $\chi^2_{\alpha;\upsilon}$ rozkładu chi-kwadrat spełniające warunek $P(\chi^2 \geq \chi^2_{\alpha;\upsilon}) = \alpha$

υ \ α	***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

f) Rozkład Pareto o parametrach η (położenia) i θ (kształtu):

Dystrybuanta: $F(x) = 1 - \left(\frac{\eta}{x}\right)^\theta$, $\eta > 0,\ \theta > 0, x \geq \eta$

Wartość oczekiwana: $E(X) = \frac{\eta\theta}{\theta-1}, \theta > 1$

g) Wartości F_{r_1,r_2} rozkładu F spełniające warunek $P(F \geq F_{r_1,r_2}) = 0.05$

r_2 \ r_1	***1***	***2***	***3***	***4***	***5***	***6***	***7***
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103
700	3.855	3.009	2.618	2.385	2.227	2.112	2.023
800	3.853	3.007	2.616	2.383	2.225	2.110	2.021
900	3.852	3.006	2.615	2.382	2.224	2.109	2.020
1000	3.851	3.005	2.614	2.381	2.223	2.108	2.019
∞	3.844	2.998	2.607	2.374	2.216	2.101	2.012

r_2 \ r_1	***8***	***9***	***10***	***11***	***12***	***13***	***14***
1	238.883	240.543	241.882	242.983	243.906	244.690	245.364
2	19.371	19.385	19.396	19.405	19.413	19.419	19.424
3	8.845	8.812	8.786	8.763	8.745	8.729	8.715
4	6.041	5.999	5.964	5.936	5.912	5.891	5.873
5	4.818	4.772	4.735	4.704	4.678	4.655	4.636
6	4.147	4.099	4.060	4.027	4.000	3.976	3.956
7	3.726	3.677	3.637	3.603	3.575	3.550	3.529
8	3.438	3.388	3.347	3.313	3.284	3.259	3.237
9	3.230	3.179	3.137	3.102	3.073	3.048	3.025
10	3.072	3.020	2.978	2.943	2.913	2.887	2.865
100	2.032	1.975	1.927	1.886	1.850	1.819	1.792
700	1.952	1.893	1.844	1.802	1.766	1.734	1.706
800	1.950	1.892	1.843	1.801	1.764	1.732	1.704
900	1.949	1.890	1.841	1.799	1.763	1.731	1.703
1000	1.948	1.889	1.840	1.798	1.762	1.730	1.702
∞	1.941	1.882	1.833	1.791	1.755	1.723	1.694

h) 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	0.567	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	0.629	0.633	0.637	0.641	0.644	0.648	0.652
0.4	0.655	0.659	0.663	0.666	0.670	0.674	0.677	0.681	0.684	0.688
0.5	0.691	0.695	0.698	0.702	0.705	0.709	0.712	0.716	0.719	0.722
0.6	0.726	0.729	0.732	0.736	0.739	0.742	0.745	0.749	0.752	0.755
0.7	0.758	0.761	0.764	0.767	0.770	0.773	0.776	0.779	0.782	0.785
0.8	0.788	0.791	0.794	0.797	0.800	0.802	0.805	0.808	0.811	0.813
0.9	0.816	0.819	0.821	0.824	0.826	0.829	0.831	0.834	0.836	0.839
1	0.841	0.844	0.846	0.848	0.851	0.853	0.855	0.858	0.860	0.862
1.1	0.864	0.867	0.869	0.871	0.873	0.875	0.877	0.879	0.881	0.883
1.2	0.885	0.887	0.889	0.891	0.893	0.894	0.896	0.898	0.900	0.901
1.3	0.903	0.905	0.907	0.908	0.910	0.911	0.913	0.915	0.916	0.918
1.4	0.919	0.921	0.922	0.924	0.925	0.926	0.928	0.929	0.931	0.932
1.5	0.933	0.934	0.936	0.937	0.938	0.939	0.941	0.942	0.943	0.944
1.6	0.945	0.946	0.947	0.948	0.949	0.951	0.952	0.953	0.954	0.954
1.7	0.955	0.956	0.957	0.958	0.959	0.960	0.961	0.962	0.962	0.963
1.8	0.964	0.965	0.966	0.966	0.967	0.968	0.969	0.969	0.970	0.971
1.9	0.971	0.972	0.973	0.973	0.974	0.974	0.975	0.976	0.976	0.977
2	0.977	0.978	0.978	0.979	0.979	0.980	0.980	0.981	0.981	0.982
2.1	0.982	0.983	0.983	0.983	0.984	0.984	0.985	0.985	0.985	0.986
2.2	0.986	0.986	0.987	0.987	0.987	0.988	0.988	0.988	0.989	0.989
2.3	0.989	0.990	0.990	0.990	0.990	0.991	0.991	0.991	0.991	0.992
2.4	0.992	0.992	0.992	0.992	0.993	0.993	0.993	0.993	0.993	0.994
2.5	0.994	0.994	0.994	0.994	0.994	0.995	0.995	0.995	0.995	0.995
2.6	0.995	0.995	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.996
2.7	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997
2.8	0.997	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998
2.9	0.998	0.998	0.998	0.998	0.998	0.998	0.998	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

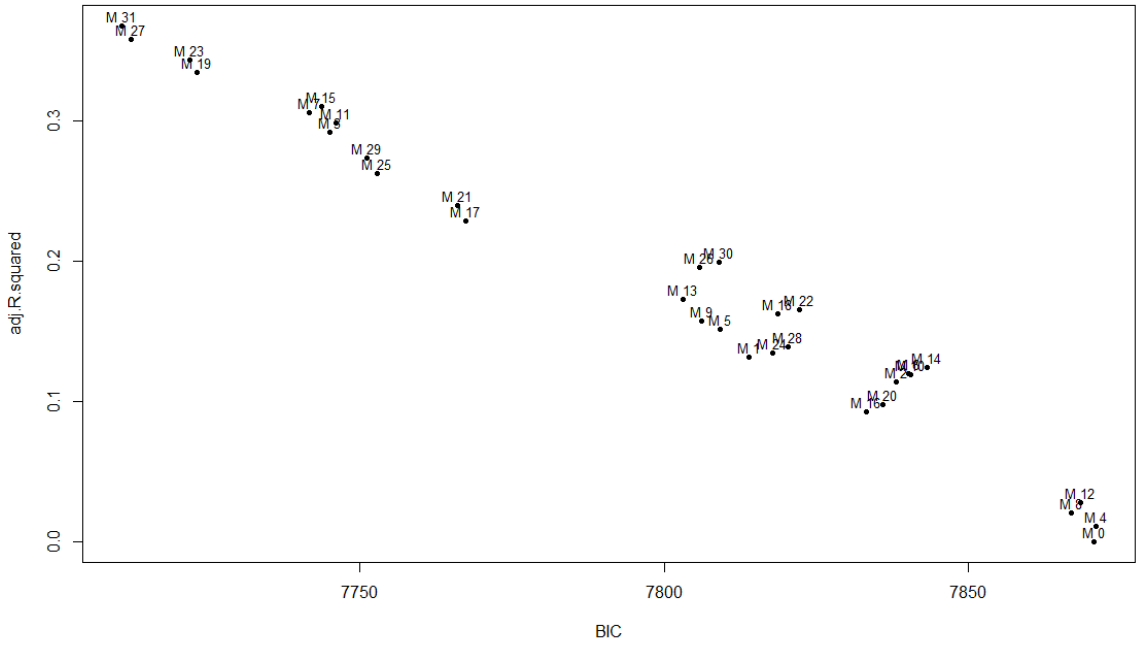
Zadanie 1.

Aktuariusz ma za zadanie opracowanie liniowego modelu regresji dla pewnej zmiennej Y. Może uwzględnić w nim maksymalnie pięć następujących regresorów (zmiennych objaśniających): X1, X2, X3, X4 i X5. W tym celu uwzględnił każdą ich kombinację i na podstawie 437 obserwacji oszacował 32 modele (łącznie z modelem zerowy). Liczbę parametrów i sumę kwadratów reszt (RSS) dla tych modeli prezentuje Tabela 1, natomiast Rys.1 przedstawia zależność między zmodyfikowanym współczynnikiem determinacji, a kryterium informacyjnym BIC.

Tabela. 1

Nazwa	Model	Liczba parametrów	RSS
M 0	Y~1 (model zerowy)	1	1651001349.0
M 1	Y~ X5	2	1430498652.0
M 2	Y~ X4	5	1449824345.0
M 4	Y~ X3	2	1155783883.0
M 8	Y~ X2	2	1629476646.0
M 16	Y~ X1	2	1395038969.0
M 3	Y~ X4+X5	6	1436648116.0
M 5	Y~ X3+X5	3	1130786778.0
M 6	Y~ X3+X4	6	1614364367.0
M 9	Y~ X2+X5	3	1385720258.0
M 10	Y~ X2+X4	6	1437640156.0
M 12	Y~ X2+X3	3	1142508102.0
M 17	Y~ X1+X5	3	1597584727.0
M 18	Y~ X1+X4	6	1356793609.0
M 20	Y~ X1+X3	3	1426450282.0
M 24	Y~ X1+X2	3	1120340291.0
M 7	Y~ X3+X4+X5	7	1494532878.0
M 11	Y~ X2+X4+X5	7	1268202827.0
M 13	Y~ X2+X3+X5	4	1367321496.0
M 14	Y~ X2+X3+X4	7	1084120247.0
M 19	Y~ X1+X4+X5	7	1483476062.0
M 21	Y~ X1+X3+X5	4	1246721785.0
M 22	Y~ X1+X3+X4	7	1359496500.0
M 25	Y~ X1+X2+X5	4	1066441791.0
M 26	Y~ X1+X2+X4	7	1422983353.0
M 28	Y~ X1+X2+X3	4	1209862151.0
M 15	Y~ X2+X3+X4+X5	8	1309604539.0
M 23	Y~ X1+X3+X4+X5	8	1043242925.0
M 27	Y~ X1+X2+X4+X5	8	1411653602.0
M 29	Y~ X1+X2+X3+X5	5	1188376486.0
M 30	Y~ X1+X2+X3+X4	8	1301217713.0
M 31	Y~ X1+X2+X3+X4+X5	9	1025185381.0

Rysunek 1.



a) Na podstawie wyników przedstawionych na rysunku 1 wskazać najlepszy i najgorszy model. Odpowiedź uzasadnić.
b) Dla najlepszego modelu obliczyć odchylenie standardowe reszt (*residual standard error*).
c) Wykorzystując odpowiedni test, sprawdzić czy wszystkie zmienne (łącznie) modelu wskazanego jako najlepszy są istotne.

Odpowiedzi:

Odp. a)
Należało wskazać model M31. Dla tego modelu aktuariusz uzyskał najwyższą wartość zmodyfikowanego współczynnika determinacji oraz najniższą wartość kryterium informacyjnym BIC.

Odp. b)
$s_\varepsilon = 1547.67$

Odp. c)
Parametry są statystycznie istotne.

Rozwiązanie:

Ad. a) Patrz odpowiedź do a).

Ad. b) Odchylenie standardowe reszt (*residual standard error*):

$$s_\varepsilon = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k}} = \sqrt{\frac{RSS}{n-k}},$$

gdzie k – liczba parametrów.

$$s_\varepsilon = \sqrt{\frac{RSS}{n-k}} = \sqrt{\frac{1025185381}{437-9}} = 1547.67$$

Ad. c)

Należało skorzystać z test F.
Wartość statystyki:

$$F = \frac{TSS - RSS}{RSS} \cdot \frac{n-k}{k-1} = \frac{R^2}{1-R^2} \cdot \frac{n-k}{k-1},$$

gdzie:
R^2- współczynnik determinacji,

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS}$$

k- liczba parametrów.

Statystyka ma rozkład Fishera-Snedecora o $r_1 = k$ *i* $r_2 = n - (k+1)$ stopniach swobody.
Stąd:

$$R^2 = 1 - \frac{1025185381}{1651001349} = 0.379$$

$$F = \frac{0.379}{1-0.379} \cdot \frac{437-9}{8} = 32.659$$

Wartość krytyczna na poziomie istotności 0.05 odczytana z tablic podanych na str. 7 wynosi 1.952 (lub 2.032).
Wniosek: Parametry są statystycznie istotne.

Zadanie 2.

W celu konstrukcji modelu predykcyjnego dla liczby szkód K_i zgłaszanych przez kierowców w ciągu roku w pewnym portfelu ubezpieczeń AC, aktuariusz może uwzględnić następujące zmienne objaśniające:

Zmienna	Opis
Wiek	Wiek kierowcy (w latach). Zmienna ilościowa
Plec	Płeć kierowcy (*Male*- mężczyzna, *Female* – kobieta)
Bon.Mal	Klasa taryfikacyjna, w której znajduje się kierowca. Zmienna jakościowa przyjmująca następujące kategorie: *0, 1, 2, 3, 4* i *5*.
Rodzaj.Sam	Rodzaj samochodu. Zmienna jakościowa przyjmująca następujące kategorie: *1, 2, 3, 4, 5* i *6*.
Obszar	Obszar zamieszkania kierowcy. Zmienna jakościowa przyjmująca następujące kategorie: *A, B, C, D, E, F, G, H, K i L*.

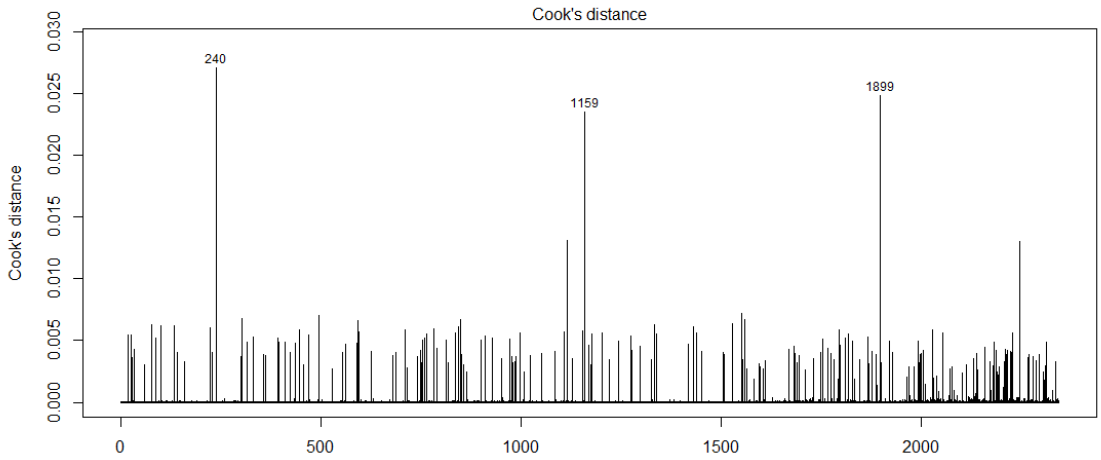
Na wstępie oszacował model z wszystkimi zmiennymi, uzyskując następujące wyniki:
- Parametry:

	Estimate	Std.Error	z Value	Pr(>|z|)
(Intercept)	-1.428	0.341	-4.193	0.000
Wiek	-0.012	0.003	-4.077	0.000
Plec Male	0.283	0.148	1.920	0.055
Bon.Mal 1	-0.427	0.212	-2.010	0.044
Bon.Mal 2	-0.508	0.210	-2.418	0.016
Bon.Mal 3	-0.808	0.235	-3.436	0.001
Bon.Mal 4	-0.999	0.256	-3.902	0.000
Bon.Mal 5	-1.124	0.256	-4.388	0.000
Rodzaj.Sam 2	0.307	0.286	1.073	0.283
Rodzaj.Sam 3	0.322	0.283	1.134	0.257
Rodzaj.Sam 4	0.125	0.293	0.425	0.671
Rodzaj.Sam 5	0.576	0.266	2.163	0.031
Rodzaj.Sam 6	0.937	0.252	3.717	0.000
Obszar B	-0.183	0.296	-0.618	0.536
Obszar C	-0.024	0.287	-0.083	0.934
Obszar D	-0.240	0.313	-0.766	0.444
Obszar E	-0.486	0.330	-1.473	0.141
Obszar F	-0.591	0.355	-1.667	0.095
Obszar G	-0.411	0.323	-1.273	0.203
Obszar H	-0.611	0.346	-1.765	0.078
Obszar K	-0.338	0.323	-1.046	0.296
Obszar L	-0.310	0.259	-1.196	0.232

- Wyniki analizy wariancji

	Df	Deviance	Resid. Df	Resid. Dev
NULL			2344	1001,47
Wiek	1	18,69	2343	982,78
Plec	1	4,14	2342	978,64
Bon.Mal	5	31,31	2337	947,33
Rodzaj.Sam	5	21,36	2332	925,97
Obszar	9	7,31	2323	918,66

- Wykres miary Cooka (*Cook's distance*)



Na podstawie podanych wyników zaproponuj aktuariuszowi model, który powinien oszacować w kolejnym etapie i wypowiedz się na temat danych, które powinien użyć do jego estymacji. **Uzasadnij swoją propozycję**. W uzasadnieniu powołaj się **między innymi** na wyniki odpowiednich testów.

Odpowiedź

W kolejnym etapie aktuariusz powinien oszacować model bez zmiennej objaśniającej *Obszar*.

W uzasadnieniu należało wskazać na:

1. Nieistotność parametrów stojących przy poszczególnych kategoriach zmiennej niezależnej *Obszar*. Wymagane było tutaj powołanie się na wyniki testu *t*-Studenta.
2. Nieistotne obniżenie dewiancji modelu ze zmienną *Obszar* w porównaniu z modelem bez tej zmiennej (z takimi samymi pozostałymi zmiennymi). Wymagane było tutaj powołanie się na test ilorazu wiarygodności. Wartość statystyki wynosi 7,31. Wartość krytyczna: $\chi^2_{0.05;9} = 16.919$.

Model powinien szacować bez uwzględnienia obserwacji o numerze: 240, 1159 i 1899.

W uzasadnieniu należało powołać się na miarę Cooka.

Uwaga! Akceptowane były także inne poprawnie uzasadnione propozycje bez zmiennej *Obszar*.

Rozwiązanie:

Patrz odpowiedź.

Zadanie 3.

W celu konstrukcji modelu predykcyjnego dla wysokości pojedynczego roszczenia Y_i (*severity model*) brano pod uwagę następujące zmienne objaśniające:

Zmienna	*Opis*
G.Ryzyka	Grupa ryzyka. Zmienna jakościowa przyjmująca następujące kategorie: *A, B, C, D, E*. Kategorie określają grupy ryzyka utworzone z uwzględnieniem wieku i doświadczenia kierowcy.
Terytorium	Terytorium zamieszkania kierowcy. Zmienna jakościowa przyjmująca następujące kategorie: *T1, T2, T3, T4, T5, T6.*

Oszacowano dwa modele M1 i M2, w których przyjęto rozkład gamma dla zmiennej objaśnianej Y_i oraz logarytmiczną funkcję wiążącą. Otrzymano następujące parametry:

Model	M1		M2	
Effect	*Estimate*	*p-Value*	*Estimate*	*p-Value*
(Intercept)	7.6528	0.0000	7.6490	0.0000
GrupaRyzyka B	0.2553	0.1110	0.2507	0.1176
GrupaRyzyka C	-0.0250	0.7403	-0.0257	0.7330
GrupaRyzyka D	-0.3037	0.0004	-0.2998	0.0005
GrupaRyzyka E	-0.2572	0.0000	-0.2633	0.0000
Terytorium T2			0.0350	0.6420
Terytorium T3			0.0578	0.4925
Terytorium T4			-0.0048	0.9459
Terytorium T5			-0.0289	0.6798
Terytorium T6			0.0022	0.9764
Parametr dyspersji	1.027517		1.025542	
	Null deviance: 2957.5 on 2531 degrees of freedom Residual deviance: 2925.9 on 2527 degrees of freedom AIC: 43604		Null deviance: 2957.5 on 2531 degrees of freedom Residual deviance: 2924.1 on 2522 degrees of freedom AIC: 43612	

oraz wyniki 10-krotnej walidacji krzyżowej (10-*fold cross validation*):

M1			M2		
RMSE	*MAE*	Resample	*RMSE*	*MAE*	Resample
2301.8	1590.4	Fold01	1931.3	1445.5	Fold01
1845.0	1432.5	Fold02	2032.0	1475.0	Fold02
2105.1	1529.7	Fold03	2077.9	1485.9	Fold03
1948.3	1455.5	Fold04	2108.6	1487.6	Fold04
2407.8	1596.1	Fold05	2066.7	1521.9	Fold05
1829.4	1407.7	Fold06	2122.5	1533.3	Fold06
2169.7	1546.1	Fold07	2075.7	1483.5	Fold07
1923.2	1487.5	Fold08	1933.2	1463.7	Fold08
1957.8	1492.9	Fold09	1952.3	1480.3	Fold09
2010.1	1449.7	Fold10	2192.4	1595.7	Fold10

Użyte w tabelach skróty *RMSE* i *MAE* oznaczają odpowiednio: średni błąd predykcji (*root mean square error*) oraz średni absolutny błąd prognoz (*mean absolute error*).

a) Uwzględniając podane wyżej informacje dotyczące parametrów oszacowanych modeli, jak również wyniki walidacji krzyżowej wybrać lepszy model (**wybór uzasadnić**).
b) Wykorzystując wybrany model (w punkcie a), wyznaczyć różnicę między prognozowaną najwyższą i najniższą wysokością pojedynczego roszczenia.

Odpowiedzi:

.....

Odp. a)

Należało wybrać model M1. W uzasadnieniu należało wskazać, że:

1. Model M1 ma mniejszą wartość kryterium informacyjnego AIC.
2. Uwzględniona w modelu M2 kolejna zmienna objaśniająca *Terytorium* nie wpłynęła na istotne obniżenie dewiancji. Tutaj należało powołać się na wyniki testu F:
   a. Wartość statystyki
   $$F = \frac{D(y;\hat{\theta}^P) - D(y;\hat{\theta}^Q)}{\hat{\phi}(q-p)} = \frac{2925.9 \ - 2924.1}{1.025542 \cdot 5} = 0.3510$$
   b. Wartość krytyczna: $F_{kr} = F_{5\,,2531} = 2.216$
   c. Wniosek: Nie ma podstaw do odrzucenia hipotezy o nieistotnej redukcji dewiancji.
3. Wyniki 10-krotnej walidacji krzyżowej są praktycznie porównywalne dla obydwu modeli. Tutaj należało powołać się na średnie wartości *RMSE* i *MAE* i zaznaczyć, że są minimalnie mniejsze dla modelu M2.

.....

Odp. b)

Najniższa wartość: 1554.797 (grupa ryzyka D).
Najwyższa wartość: 2719.219 (grupa ryzyka B).
Różnica: 1164.425

Rozwiązanie:

Ad. a) Patrz odpowiedź do a)

Ad. b)
Najniższa wartość będzie dla grupy ryzyka D:
$\hat{y}_D = \exp(7.6528 - 0.3037) = 1554.797$

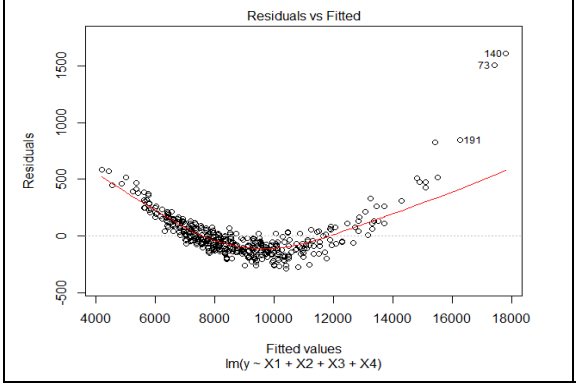
Najwyższa wartość będzie dla grupy ryzyka B:
$\hat{y}_D = \exp(7.6528 + 0.2507) = 2719.219$

Zadanie 4.

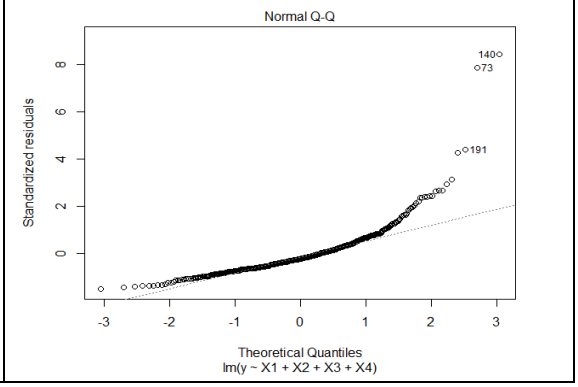
Wykorzystując klasyczną metodę najmniejszych kwadratów, oszacowano model regresji liniowej dla zmiennej *Y*. Uwzględniono w nim cztery zmienne objaśniające X1, X2, X3 i X4. Przeprowadzono analizę reszt tego modelu, uzyskując następujące wykresy:

(i) diagnostyczne:

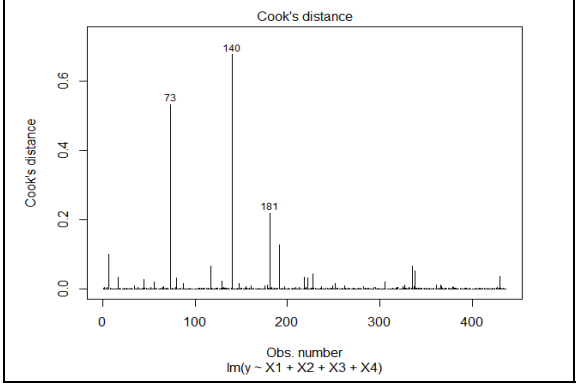
Rys. A



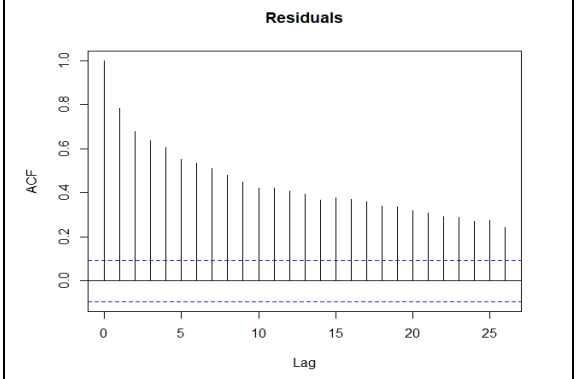
Rys. B



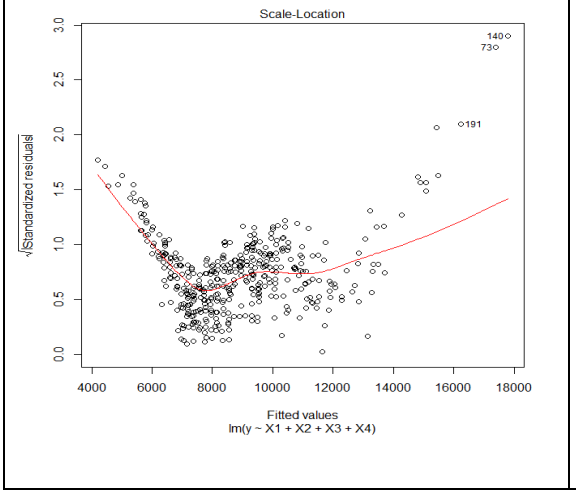
Rys. C



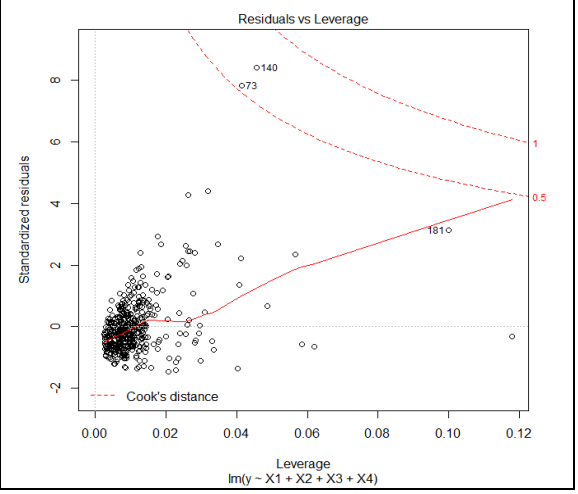
Rys. D



Rys. E

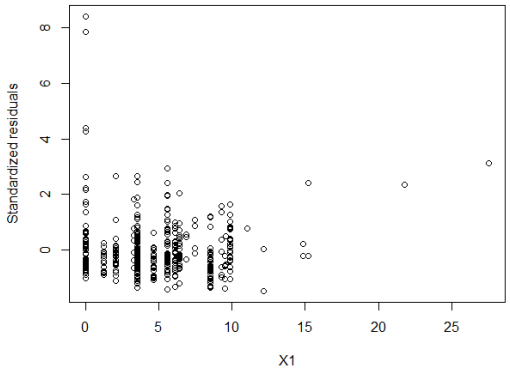


Rys. F

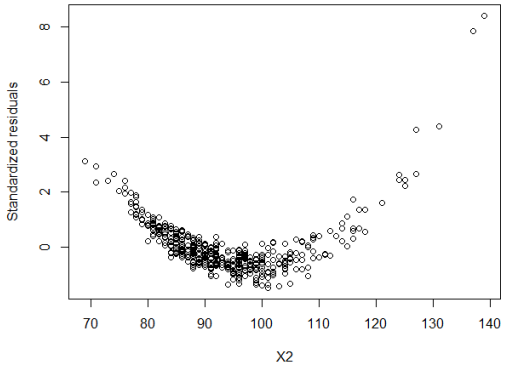


(ii) zależności standaryzowanych reszt od wartości zmiennych niezależnych:

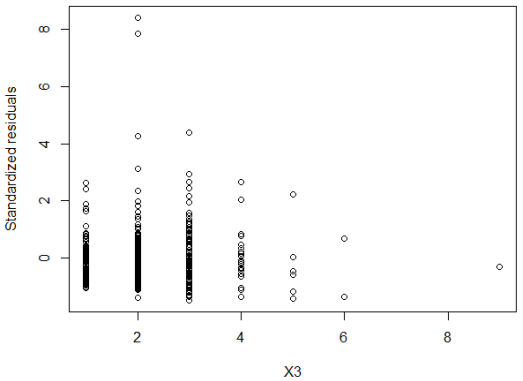
Rys. G



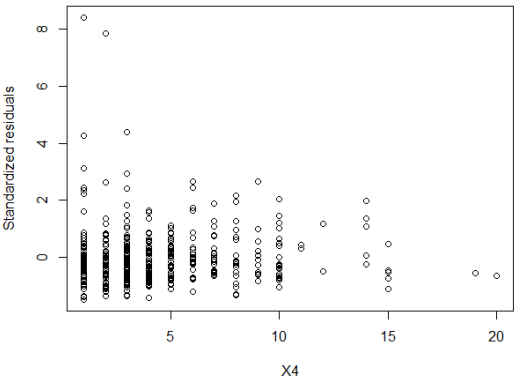
Rys. H



Rys. I



Rys. J



W oparciu o zaprezentowane wykresy omówić strukturę stochastyczną reszt modelu. Czy spełnia ona założenia klasycznej metody najmniejszych kwadratów. W odpowiedzi należy wskazać odpowiednie założenia i napisać (**powołując się na konkretny wykres**), czy **mogą** być spełnione (to czy są, czy nie są spełnione ostatecznie rozstrzyga odpowiedni test). W przypadku odpowiedzi negatywnej zaproponować model, którego reszty mogą spełniać więcej założeń. Proponując konkretną postać modelu należy wskazać, które aspekty struktury stochastycznej modelu mogą ulec poprawie.

Odpowiedź

Powołując się na odpowiednie wykresy diagnostyczne (i), należało wskazać, że mogą wystąpić problemy z:

1. Normalnością reszt.
2. Jednorodnością wariancji reszt.
3. Niezależnością reszt.

Na podstawi wykresów zależności standaryzowanych reszt od wartości zmiennych niezależnych (ii) należało zaproponować model, w którym dodatkowo uwzględnia się kwadraty wartości zmiennej X2.
Uwaga! Akceptowane były także inne poprawnie uzasadnione propozycje.

Rozwiązanie:

Patrz odpowiedź.

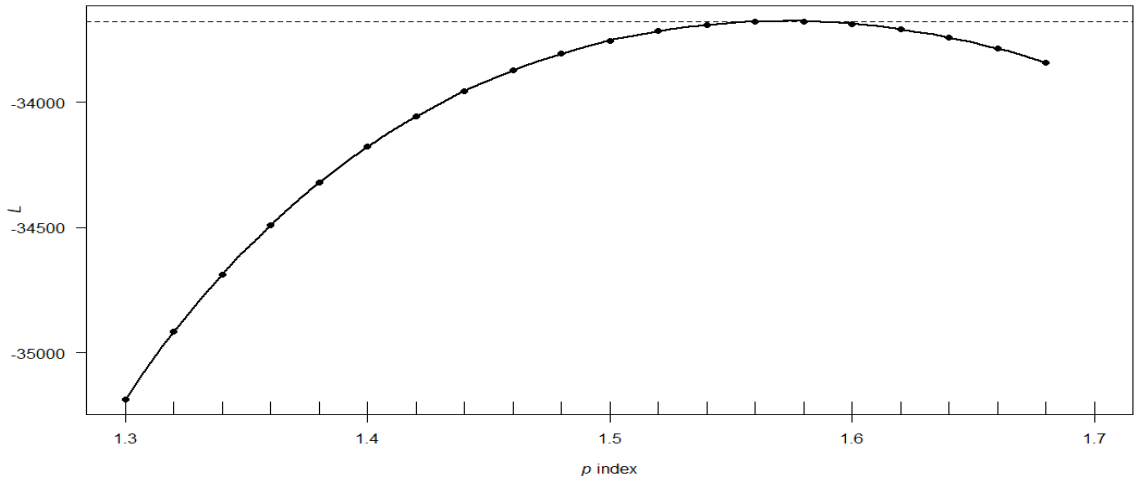
Zadanie 5.

Uzyskano następujące wyniki jednomodelowej taryfikacji z wykorzystaniem uogólnionego modelu liniowego (GLM):

Zmienna zależna Y_i:	składka czysta (*pure premium*) – całkowite roszczenie dla *i*-tej polisy przez jej ekspozycję
Rozkład zmiennej zależnej:	Tweedie
Parametr dyspersji (ϕ):	100
Funkcja wiążąca:	logarytm naturalny (*informacja podana na początku egzaminu*)

Parametr:	Estimate	Std.Error
(Intercept)	4.57	0.12
Grupa ryzyka:		
GR.1	Referencyjna	
GR.2	0.04	0.13
GR.3	0.38	0.23
GR.4	0.77	0.44
GR.5	-0.08	0.42
GR.6	1.57	0.42
GR.7	0.19	1.08
GR.8	1.55	0.26
GR.9	1.20	0.38
Obszar zamieszkania:		
Ob.1	Referencyjny	
Ob.2	0.11	0.16
Ob.3	0.41	0.17
Ob.4	0.33	0.15
Ob.5	0.55	0.15
Ob.6	0.85	0.16

Wartość logarytmu funkcji wiarygodności modelu w zależności od parametru *p*:



a) Wyznaczyć najwyższą ($\hat{Y}_{max}$) i najniższą ($\hat{Y}_{min}$) składkę czystą otrzymaną z wykorzystaniem tego modelu. Przyjąć optymalną wartość parametru (indeksu) p.
b) Dla składek $\hat{Y}_{max}$ i $\hat{Y}_{min}$ obliczyć wariancję i odchylenie standardowe.

Odpowiedzi:

.....

Odp. a)

$\hat{Y}_{max} = 1085.721$ (dla grupa ryzyka: GR.6 i obszaru zamieszkania: Ob.6)
$\hat{Y}_{min} = 89.121$ (dla grupa ryzyka: GR.5 i obszaru zamieszkania: Ob.1)

.....

Odp. b)

Dla $\hat{Y}_{max}$ wariancja wynosi: 5441526.499, a odchylenie standardowe 2332.708
Dla $\hat{Y}_{min}$ wariancja wynosi: 110146.898, a odchylenie standardowe 331.884

Rozwiązanie:

Jako optymalną wartość parametru (indeksu) p należało przyjąć 1.56 (**uwzględniane były także parametry z przedziału od 1.54 do 1.58).**

Ad. a)

$\hat{Y}_{max} = \exp(4.57 + 1.57 + 0.85) = 1085.721$
$\hat{Y}_{min} = \exp(4.57 - 0.88) = 89.121$

Ad. b)
Dla rozkładu Tweedie funkcja wariancji ma postać: $\vartheta(\mu) = \mu^p$.
Zatem:
- dla $\hat{Y}_{max}$ wariancja wynosi: $100 \cdot 1085.721^{1.56} = 5441526.499$, a odchylenie standardowe 2332.708
- dla $\hat{Y}_{min}$ wariancja wynosi: $100 \cdot 89.121^{1.56} = 110146.898$, a odchylenie standardowe 331.884

Zadanie 6.

Roczne straty

- dla pierwszej linii biznesu (LOB1) są modelowane za pomocą zmiennej losowej X o rozkładzie Pareto z parametrami $\eta = 100$ i $\theta = 2$,
- dla drugiej linii biznesu (LOB2) - zmiennej losowej Y o rozkładzie Pareto z parametrami $\eta = 200$ i $\theta = 2$.

Obliczyć prawdopodobieństwo łączne, że starty z LOB1 są większe od mediany X i mniejsze od wartości zagrożonej $VaR_{0.95}(X)$ oraz starty z LOB2 są większe od mediany Y i mniejsze od wartości zagrożonej $VaR_{0.95}(Y)$, czyli

$$P(Med(X) < X < VaR_{0.95}(\mathrm{X}), Med(Y) < Y < VaR_{0.95}(Y)\ ),$$

przy założeniu, że

a) zmienne X i Y są niezależne,
b) struktura zależności między zmiennymi X i Y jest modelowana za pomocą kopuli Gumbela-Hougaarda o generatorze $\phi(t) = (-\ln t\ )^2$.

Odpowiedzi:

.....

Odp. a)

$$P(Med(X) < X < VaR_{0.95}(\mathrm{X}), Med(Y) < Y < VaR_{0.95}(Y)\ ) = 0.2025$$

.....

Odp. b)

$$P(Med(X) < X < VaR_{0.95}(\mathrm{X}), Med(Y) < Y < VaR_{0.95}(Y)\ ) = 0.307108$$

Rozwiązanie:

Niech: $a = Med(X)$, $b = VaR_{0.95}(\mathrm{X})$, $c = Med(Y)$ i $d = VaR_{0.95}(Y)$,
wówczas

$$P(Med(X) < X < VaR_{0.95}(\mathrm{X}), Med(Y) < Y < VaR_{0.95}(Y)\ ) = \\ = F_{(X,Y)}(b,d) - F_{(X,Y)}(a,d) - F_{(X,Y)}(b,c) + F_{(X,Y)}(a,c)$$

Dalej $F_X(x)$, $F_Y(y)$ oznaczają dystrybuanty zmiennych odpowiednio X i Y.

1. Jeżeli zmienne X i Y są niezależne, to $F_{(X,Y)}(x,y) = F_X(x) \cdot F_Y(y)$.
   Otrzymujemy
   $F_{(X,Y)}(b,d) = F_X(b) \cdot F_Y(d) = F_X\left(F_X^{-1}(0{,}95)\right) \cdot F_Y\left(F_Y^{-1}(0{,}95)\right) = 0{,}95 \cdot 0{,}95 = \\ = 0.9025$
   $F_{(X,Y)}(a,d) = F_X(a) \cdot F_Y(d) = F_X\left(F_X^{-1}(0{,}50)\right) \cdot F_Y\left(F_Y^{-1}(0{,}95)\right) = 0{,}50 \cdot 0{,}95 = \\ = 0.475$
   $F_{(X,Y)}(b,c) = F_X(b) \cdot F_Y(c) = F_X\left(F_X^{-1}(0{,}95)\right) \cdot F_Y\left(F_Y^{-1}(0{,}50)\right) = 0{,}95 \cdot 0{,}50 = \\ = 0.475$
   $F_{(X,Y)}(a,c) = F_X(a) \cdot F_Y(c) = F_X\left(F_X^{-1}(0{,}50)\right) \cdot F_Y\left(F_Y^{-1}(0{,}50)\right) = 0{,}50 \cdot 0{,}50 = \\ = 0.25$
   Stąd:
   $$P(Med(X) < X < VaR_{0.95}(\mathrm{X}), Med(Y) < Y < VaR_{0.95}(Y)\ ) = 0.2025$$

2. Jeżeli zależność między zmiennymi X i Y jest modelowana za pomocą kopuli C, to $F_{(X,Y)}(x,y) = C(F_X(x), F_Y(y))$.

Otrzymujemy:

$F_{(X,Y)}(b,d) = C\left(F_X\left(F_X^{-1}(0{,}95)\right), F_Y\left(F_Y^{-1}(0{,}50)\right)\right) = C(0.95\,, 0.95)$

W zadaniu C jest kopulą Archimedesa, zatem: $C = \phi^{-1}(\phi(u) + \phi(v))$.

Stąd:

$\phi(0{,}95) = (-\ln(0.95))^2 = 0.002631$

$\phi^{-1}(z) = \exp(-\sqrt{z})$

$C(0.95\,, 0.95) = exp\left(-\sqrt{2 \cdot 0.002631}\right) = 0{,}9300$

$F_{(X,Y)}(a,d) = C\left(F_X\left(F_X^{-1}(0.50)\right), F_Y\left(F_Y^{-1}(0.95)\right)\right) = C(0.50\,, 0.95)$

W zadaniu C jest kopulą Archimedesa, zatem: $C = \phi^{-1}(\phi(u) + \phi(v))$.

Stąd:

$\phi(0{,}95) = (-\ln(0.95))^2 = 0.002631$

$\phi(0{,}50) = (-\ln(0.50))^2 = 0.480453$

$\phi^{-1}(z) = \exp(-\sqrt{z})$

$C(0.50\,, 0.95) = exp\left(-\sqrt{0.480453 + 0.002631}\right) = 0.499053$

$F_{(X,Y)}(b,c) = C\left(F_X\left(F_X^{-1}(0.95)\right), F_Y\left(F_Y^{-1}(0.50)\right)\right) = C(0.95\,, 0.50)$

W zadaniu C jest kopulą Archimedesa, zatem: $C = \phi^{-1}(\phi(u) + \phi(v))$.

Stąd:

$\phi(0{,}95) = (-\ln(0.95))^2 = 0.002631$

$\phi(0{,}50) = (-\ln(0.50))^2 = 0.480453$

$\phi^{-1}(z) = \exp(-\sqrt{z})$

$C(0.50\,, 0.95) = exp\left(-\sqrt{0.480453 + 0.002631}\right) = 0.499053$

$F_{(X,Y)}(a,c) = C\left(F_X\left(F_X^{-1}(0.50)\right), F_Y\left(F_Y^{-1}(0.50)\right)\right) = C(0.50\,, 0.50)$

W zadaniu C jest kopulą Archimedesa, zatem: $C = \phi^{-1}(\phi(u) + \phi(v))$.

Stąd:

$\phi(0{,}50) = (-\ln(0.50))^2 = 0.480453$

$\phi^{-1}(z) = \exp(-\sqrt{z})$

$C(0.50\,, 0.50) = exp\left(-\sqrt{2 \cdot 0.480453}\right) = 0.375214$

Stąd:

$P(Med(X) < X < VaR_{0.95}(\mathrm{X}), Med(Y) < Y < VaR_{0.95}(Y)\,) = 0.307108$

Zadanie 7.

a) Na czym polega uczenie z nadzorem i uczenie bez nadzoru.
b) Wskazać i krótko scharakteryzować co najmniej dwie metody uczenia z nadzorem i co najmniej dwie metody uczenia bez nadzoru.

Odpowiedzi:

.....

Odp. a)

W odpowiedzi należało krótko omówić uczenie z nadzorem (*supervised learning*) i uczenie bez nadzoru (*unsupervised learning*). Celem uczenia z nadzorem jest znalezienie funkcji odwzorowującej dane wejściowe na dane wyjściowe na podstawie dostarczonych przez człowieka przykładowych par wejścia-wyjścia. Natomiast uczenie bez nadzoru polega na znalezieniu funkcji opisującej strukturę danych nieetykietowanych, czyli takich które nie zostały wcześniej sklasyfikowane lub skategoryzowane.

Odp. a)

Należało wskazać i krótko omówić:

1. dwie metody uczenia z nadzorem, np. regresja, klasyfikacja (*classification*),
2. dwie metody uczenia bez nadzoru, np. analiza skupień (*clustering*), redukcja wymiaru (*dimensionality reduction*).

Rozwiązanie:

Ad. a)
Patrz odpowiedź do a).

Ad. b)
Patrz odpowiedź do b).

Zadanie 8.

Wystąpienie oszustw w zgłaszanych roszczeniach w pewnym zakładzie ubezpieczeń modelowano za pomocą regresji logistycznej. Przyjęto, że zmienna zależna Y_i przyjmuje dwie wartości:
$Y_i = 1$, gdy wystąpiło oszustwo w i-tym roszczeniu,
$Y_i = 0$, gdy nie wystąpiło oszustwo w i-tym roszczeniu.
Wykorzystując ten model, dla reprezentatywnej próby liczącej 10 roszczeń uzyskano następujące prognozy prawdopodobieństwa wystąpienia oszustwa:

Nr roszczenia	Zaobserwowana (faktyczna) wartość zmiennej Y_i	Prognoza prawdopodobieństwa wystąpienia oszustwa
1	1	0.40
2	1	0.25
3	0	0.27
4	0	0.73
5	1	0.90
6	0	0.34
7	0	0.25
8	1	0.80
9	1	0.45
10	0	0.34

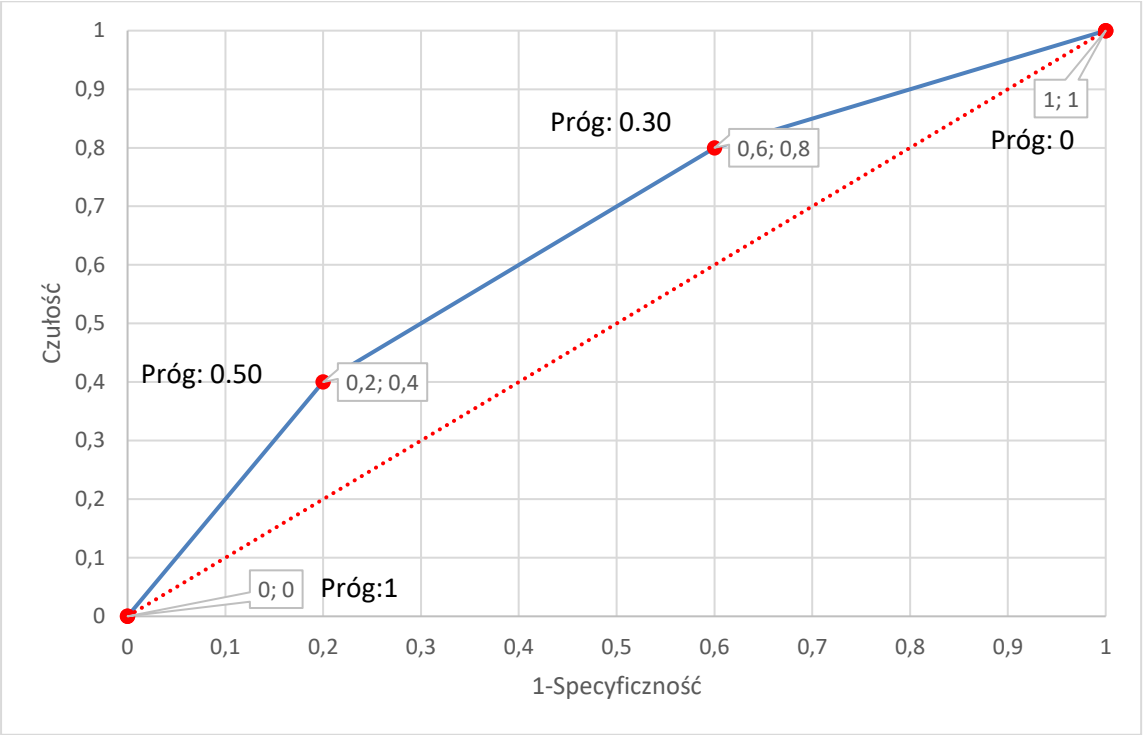
W drugiej kolumnie powyższej tabeli podano zaobserwowane (faktyczne) wartości zmiennej Y_i.

a) Uwzględniając podane wyniki narysować krzywą ROC na podstawie dwóch wartości progowych (dwóch punktów odcięcia): 0.30 i 0.50. Opisać osie współrzędnych oraz dla każdego punktu na krzywej podać jego współrzędne i opowiadającą mu wartość progową.
b) W przypadku zgłoszenia roszczenia, za pomocą opracowanego modelu sprawdza się możliwość wystąpienia oszustwa. Gdy prognoza jest pozytywna (tzn. prognozuje się, że zmienna $Y_i = 1$) wdrażana jest kontrola, której koszty nie zależą od wysokości roszczenia. Która wartość progowa 0.30, czy 0.50 jest bardziej adekwatna dla linii biznesu, charakteryzującej się niskim prawdopodobieństwem wystąpienia roszczenia i wysoką jego wartością? Odpowiedź uzasadnić.

Odpowiedzi:

.....

Odp. a)



.....

Odp. b)

Należało wybrać próg 0.3. Taka wartość progowa oznacza częstsze kontrole, których koszty mogą okazać się jednak niższe w porównaniu z kosztami niesłusznie wypłaconych wysokich odszkodowań w tej linii biznesu.

Rozwiązanie:

Ad. a)

- Tabela trafności prognoz (macierz błędu, *ang. confusion matrix*):

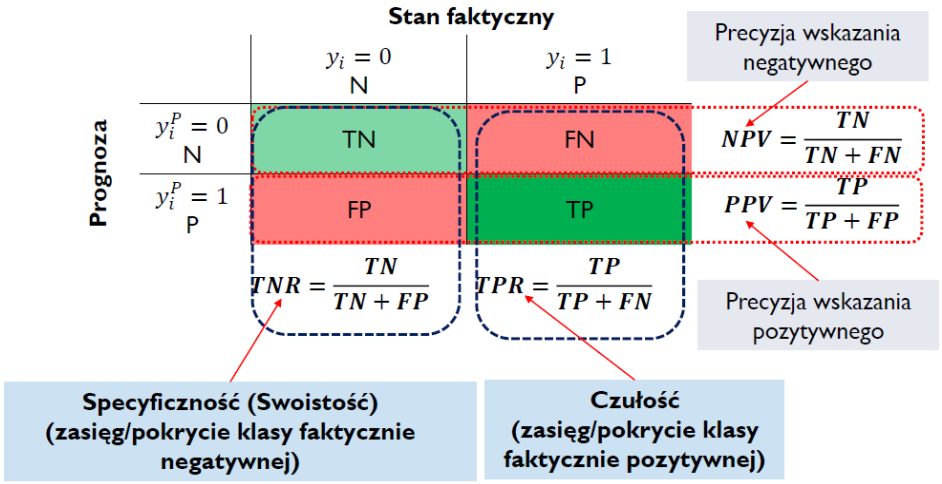


Tabela trafności dla progu 0.5

Prognozy	Faktyczne		Razem
	N	P	
N	4	3	7
P	1	2	3
Razem	5	5	10

Specyficzność: 0.8
Czułość: 0.4

Tabela trafności dla progu 0.3

Prognozy	Faktyczne		Razem
	N	P	
N	2	1	3
P	3	4	7
Razem	5	5	10

Specyficzność: 0.4
Czułość: 0.8

Ad. b)

Patrz odpowiedź do b).

Zadanie 9.

Przedłużenie umowy ubezpieczenia AC modelowano z wykorzystaniem uogólnionych modeli liniowych (GLM). Przyjęto, że zmienna zależna Y_i przyjmuje dwie wartości:
$Y_i = 1$, gdy kierowca przedłuży umowę na kolejny rok
$Y_i = 0$, gdy nie przedłuży.
Przyjęto, że zmienna Y_i ma rozkład Bernoulliego (zero-jedynkowy) oraz uwzględniono następujące zmienne objaśniające:

- *plec:* Płeć (K – kobieta, M – mężczyzna)
- *zamieszkanie:* Miejsce zamieszkania (Miasto – miasto, Wies – wieś)
- *uzytkowanie:* Użytkowanie samochodu (Sluzb – do celów służbowych, Pryw – do celów prywatnych)
- *wiek:* Wiek kierowcy w latach.
- *lojalnosc:* Liczba lat, w których kierowca był klientem zakładu ubezpieczeń.

Metodą największej wiarygodności oszacowano następujące trzy modele:

- Model 1 (**M1**), z kanoniczną funkcją wiążącą,
- Model 2 (**M2**), z funkcją wiążącą *probit*,
- Model 3 (**M3**), z funkcją wiążącą postaci: $\eta = \ln(-\ln(1-\mu))$ (funkcja *cloglog*).

Uzyskano następujące oszacowania:

Coefficients:

	M1	**M2**	**M3**
(Intercept)	2.00	1.00	1.00
plec M	-0.94	-0.55	-0.6
zamieszkanie Wies	-1.17	-0.69	-0.77
uzytkowanie Sluzb	-1.65	-0.72	-1.56
wiek	-0.06	-0.04	-0.05
lojalnosc	0.14	0.08	0.10

a) Co to jest kanoniczna funkcja wiążąca?
b) Jakie są prognozy przedłużenia umowy (zmiennej Y_i), uzyskane za pomocą oszacowanych modeli, dla mieszkającego na wsi, 43-letniego mężczyzny użytkującego samochód do celów prywatnych, który był klientem zakładu przez 18 lat? Przyjąć wartość progową (punkt odcięcia) równą 0.40.

Odpowiedzi:

.....

Odp. a)

Funkcję $g(\cdot)$ nazywamy kanoniczną funkcją wiążącą, gdy $\theta_i = g(\mu_i)$, gdzie θ_i oznacza parametr kanoniczny rozkładu zmiennej losowej Y_i (należącego do wykładniczej rodziny rozkładów).

Odp. b)

Model 1 (**M1**): kierowca przedłuży umowę na kolejny rok
Model 2 (**M2**): kierowca nie przedłuży umowę na kolejny rok
Model 3 (**M3**): kierowca nie przedłuży umowę na kolejny rok

Rozwiązanie:

Ad. a) Patrz odpowiedź a).

Ad. b) Dla rozkładu Bernoulliego kanoniczną funkcją wiążącą jest *logit:*

$$g(\mu) = ln\frac{\mu}{1-\mu}$$

Model 1 (**M1**):
- predyktor liniowy: -0.17
- prognoza prawdopodobieństwa: $\hat{p} = \frac{\exp(-0.17)}{1+\exp(-0.17)} = 0.4576020$
- prognoza zmiennej Y_i: kierowca przedłuży umowę na kolejny rok

Model 2 (**M2**):
- predyktor liniowy: -0.52
- prognoza prawdopodobieństwa: $\hat{p} = \Phi(-0.52) = 1 - \Phi(0.52) = 1 - 0.698 = 0.302$
- prognoza zmiennej Y_i: kierowca nie przedłuży umowę na kolejny rok

Model 3 (**M3**):
- predyktor liniowy: -0.72
- prognoza prawdopodobieństwa: : $\hat{p} = 1 - \exp(-\exp(-0.72)) = 0.3854$
- prognoza zmiennej Y_i: kierowca nie przedłuży umowę na kolejny rok

Zadanie 10.

Wykorzystując klasyczną metodę najmniejszych kwadratów, oszacowano dwa modele regresji liniowej dla zmiennej Y. W pierwszym (**M1**) jako zmienne objaśniające przyjęto X1 i X2, a w drugim (**M2**) - X1 i X3. Uzyskano następujące wyniki:

M1	
	Estimate
(Intercept)	8.56
X1	-0.21
X2	0.02

Residual standard error: 1.71 on 434 degrees of freedom

$$(X'X)^{-1}= \begin{vmatrix} 0.0147411 & -0.0008957 & -0.0001547 \\ -0.0008957 & 0.0001865 & -0.0000004 \\ -0.0001547 & -0.0000004 & 0.0000030 \end{vmatrix}$$

Współczynnik korelacji liniowej Pearsona między zmiennymi X1 i X2 jest równy 0.85

M2	
	Estimate
(Intercept)	9.70
X1	-0.35
X3	0.07

Residual standard error: 1.81 on 434 degrees of freedom

$$(X'X)^{-1}= \begin{vmatrix} 0.0067815 & -0.0010375 & 0.0000619 \\ -0.0010375 & 0.0025877 & -0.0012092 \\ 0.0000619 & -0.0012092 & 0.0006089 \end{vmatrix}$$

Współczynnik korelacji liniowej Pearsona między zmiennymi X1 i X3 jest równy 0.96

a) Wybrać model, w którym nie występuje problem współliniowości.
b) Wykorzystując wybrany model, wyznaczyć prognozę dla zmiennej Y w przypadku, gdy X1 =10, X2=100, X3=10. Oszacować jej błąd (***wystarczy wstawić odpowiednie wielkości do wzoru***).

Odpowiedzi:

.....

Odp. a)
Model **M1**

.....

Odp. b)

Prognoza: 8.46

Błąd prognozy: 1.721

Rozwiązanie:

Ad. a)

Współczynnik inflacji wariancji

$$VIF_j = \frac{1}{1 - R_j^2}$$

Współczynnik determinacji modelu regresji liniowej X1 względem X2 wynosi: $0.85^2 = 0.7225$.
Stąd VIF dla **M1** jest równy:

$$VIF_{M1} = \frac{1}{1 - 0.7225} = 3.604$$

Współczynnik determinacji modelu regresji liniowej X1 względem X3 wynosi: $0.96^2 = 0.9216$.
Stąd VIF dla **M2** jest równy:

$$VIF_{M2} = \frac{1}{1 - 0.9216} = 12.755$$

Ad. b)

Prognoza: $y^P = 8.56 - 0.21 \cdot 10 + 0.02 \cdot 100 = 8.46$

Błąd prognozy: $s_D = \sqrt{s_\varepsilon^2 (1 + \boldsymbol{x}_z (X'X)^{-1} \boldsymbol{x}_z')}$

$$s_D = \sqrt{1.71^2 \cdot (1 + [1 \quad 10 \quad 100] \cdot (X'X)^{-1} \cdot \begin{bmatrix} 1 \\ 10 \\ 100 \end{bmatrix}} = 1.721$$

Sesja egzaminacyjna w dniu 3 marca 2020 r.

Modelowanie

Arkusz ocen

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