Komisja Egzaminacyjna dla Aktuariuszy

LXXXV Egzamin dla Aktuariuszy

Sesja egzaminacyjna w dniu 9 czerwca 2022 r.

Modelowanie

Imię i nazwisko osoby egzaminowanej: ....

Czas trwania egzaminu: 120 minut

Uwagi

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

b) 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

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	0.567	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.

Wysokość pojedynczego roszczenia Y_i w pewnym portfelu ubezpieczeń AC modelowano z uwzględnieniem dwóch zmiennych objaśniających:

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

Oszacowano uogólniony model liniowy, w którym uwzględniono powyższe zmienne objaśniające i założono rozkład gamma dla Y_i. Przyjęto kanoniczną funkcję łączącą. Uzyskano następujące wyniki:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.460e-04	1.439e-05	51.825	< 2e-16 ***
*CarAg*e	6.804e-06	1.498e-06	4.543	5.59e-06 ***
*Brand*B	-5.203e-06	3.562e-05	-0.146	0.88388
*Brand*C	8.323e-06	3.666e-05	0.227	0.82037
*Brand*D	8.680e-05	5.756e-05	1.508	0.13152
*Brand*E	-5.828e-06	4.490e-05	-0.130	0.89673
*Brand*F	4.322e-06	2.462e-05	0.176	0.86067
*Brand*G	7.993e-05	7.099e-05	1.126	0.26025
*CarAge:Brand*B	2.049e-07	4.099e-06	0.050	0.96013
*CarAge:Brand*C	1.148e-06	4.173e-06	0.275	0.78323
*CarAge:Brand*D	-5.024e-06	6.650e-06	-0.755	0.44999
*CarAge:Brand*E	-4.010e-06	4.372e-06	-0.917	0.35905
*CarAge:Brand*F	-1.147e-05	4.318e-06	-2.656	0.00791 **
*CarAge:Brand*G	-1.021e-05	7.537e-06	-1.355	0.17549

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for Gamma family taken to be **0.7826157**)

Null deviance: 11904 on 15866 degrees of freedom
Residual deviance: 11861 on 15853 degrees of freedom
AIC: 257036

a) (***1p.***) Podaj definicję kanonicznej funkcji łączącej (linku kanonicznego)?
b) (***2p.***) Jaką postać ma kanoniczna funkcja łącząca w uogólnionym modelu liniowym, w którym zmienna objaśniana ma rozkład gamma.
Funkcja gęstości i wartość oczekiwana dla rozkładu gamma są odpowiednio równe: $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} exp(-\beta x),\ x > 0;\ \mu = \frac{\alpha}{\beta}$.
c) (***1p.***) Podaj postać wiersza w macierzy obserwacji (*design matrix, model matrix, regressor matrix*), który odpowiada 10-letniemu samochodowi marki F (wektora obserwacji dla tego samochodu).
d) (***1p.***) Wykorzystując oszacowany model, wyznacz wariancję wysokości pojedynczej szkody dla tego samochodu (10-letni samochód marki F).

Odpowiedzi

....

Odp. a)

Funkcję $g(\cdot)$ nazywamy linkiem kanonicznym (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), natomiast $\mu_i = E(Y_i)$.

....

Odp. b)

Przyjmijmy następującą parametryzację wykładniczej rodziny rozkładów:

$$f(y_i;\theta_i,\phi) = exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right),$$

gdzie: θ_i - parametr kanoniczny, ϕ - parametr dyspersji.
Wiadomo, że $\mu_i = b'(\theta_i)$, czyli $\theta_i = {b'}^{-1}(\mu_i)$. Dla rozkładu gamma $b(\theta_i) = -\ln(-\theta_i)$, skąd otrzymujemy link kanoniczny postaci $g(\mu_i) = -\frac{1}{\mu_i}$.

....

Odp. c)

$$\boldsymbol{x}_i = [1\ \ 10\ \ 0\ \ 0\ \ 0\ \ 0\ \ 1\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 10\ \ 0]$$

....

Odp. d)

$$Var(Y_i) = 1580594{,}11$$

Rozwiązanie:

$\boldsymbol{x}_i\boldsymbol{\beta}^T = 0{,}00070366$, stąd $\mu_i = 1421.1369$
$Var(Y_i) = \phi \cdot {\mu_i}^2 = 0.7826157 \cdot 1421.1369^2 = 1580594{,}11$

Zadanie 2.

Dla pewnego portfela ubezpieczeń badano zależność rocznej liczby szkód (zmienna K_i) od wieku ubezpieczonego wyrażonego w latach (zmienna ilościowa *wiek*) oraz od tego, czy szkoda wystąpiła w poprzednim roku (zmienna jakościowa *szkoda.poprzedni.rok*, przyjmująca dwie wartości: „TAK”- gdy szkoda wystąpiła oraz „NIE”- gdy szkoda nie wystąpiła. Oszacowano dwa modele regresji Poissona z kanonicznymi funkcjami łączącymi (linkami kanonicznymi). W obydwu modelach jako zmienną offsetową uwzględniono czas ekspozycji (w latach) na skali kanonicznej. Uzyskano następujące wyniki:

<u>Model **M1:**</u>

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.410969	0.093913	-15.024	< 2e-16 ***
wiek	-0.009919	0.002089	-4.749	2.04e-06 ***
*szkoda.poprzedni.rok*TAK	0.088382	0.044637	1.980	0.0477 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 9881.4 on 24455 degrees of freedom
Residual deviance: 9856.0 on 24453 degrees of freedom
AIC: 13740

<u>Model **M2**</u> (zmienna $wiek.kw = wiek^2$)**:**

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.2613798	0.2844895	-0.919	0.3582
wiek	-0.0633078	0.0126683	-4.997	5.81e-07 ***
wiek.kw	0.0005816	0.0001354	4.294	1.75e-05 ***
*szkoda.poprzedni.rok*TAK	0.0944986	0.0446955	2.114	0.0345 *

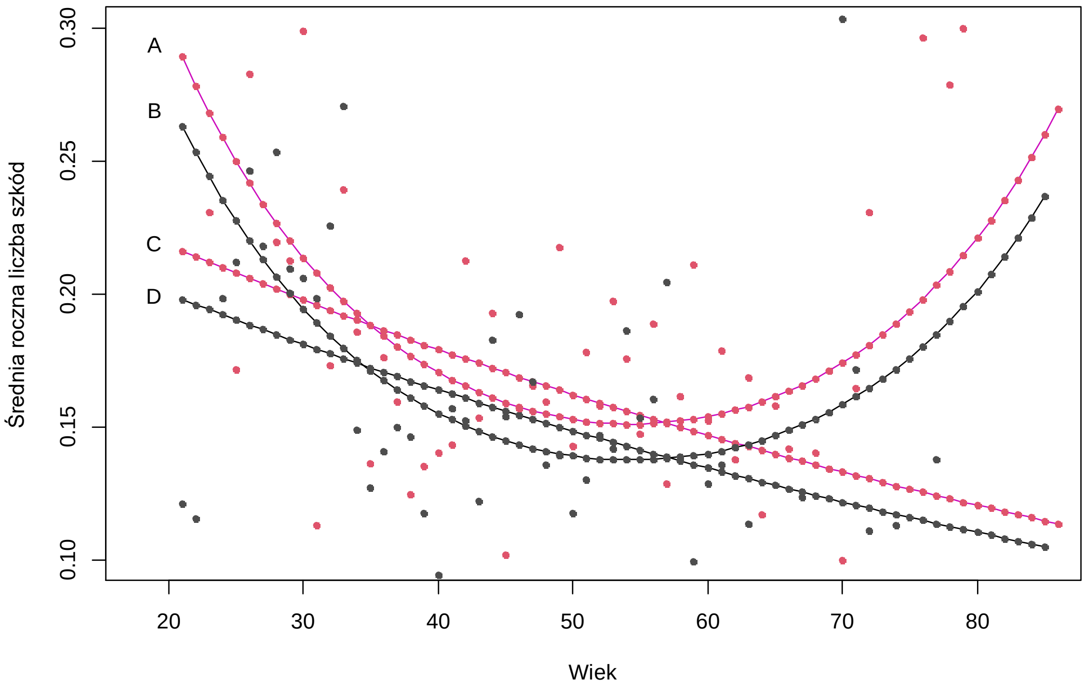
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 9881.4 on 24455 degrees of freedom
Residual deviance: 9839.1 on 24452 degrees of freedom
AIC: 13725

Na poniższym rysunku (Rys. 2.1) przedstawiono zależność średniej rocznej liczby szkód od wieku i wystąpienia szkody w poprzednim roku, otrzymaną na podstawie danych rzeczywistych i oszacowanych modeli M1 i M2. Wystąpienie lub niewystąpienie szkody rozróżniono kolorem.

Rysunek 2.1



a) (***2p.***) Opisz składowe wykresu przedstawionego na rysunku 2.1, w szczególności krzywe A, B, C, D i punkty „luźno rozrzucone". Który kolor oznacza wystąpienie, a który niewystąpienie szkody w poprzednim roku? **Odpowiedzi odpowiednio uzasadnij**!

b) (***1p.***) Wskaż, który z tych dwóch modeli jest lepszy. Wybór uzasadnij w oparciu o przedstawione wyniki oszacowań modeli i wykres przedstawiony na rysunku 2.1.

c) (***2p.***) Do zbioru uczącego wykorzystanego do oszacowania obydwu modeli należy 30-sto letni ubezpieczony z sześciomiesięczną ekspozycją na ryzyko (*ekspozycja =0.5*), w czasie której zgłosił jedną szkodę. W jego przypadku nie zanotowano szkody w poprzednim roku. Dla obydwu modeli wyznacz resztę Pearsona odpowiadającą tej obserwacji.

Odpowiedzi:

....

Odp. a)

Kolorem czerwonym oznaczono średnią roczną liczbę szkód dla kierowców, w przypadku których zanotowano szkodę w poprzednim roku. Wskazuje na to dodatnia wartość parametru przy zmiennej *szkoda.poprzedni.rok*TAK. Opis krzywych:

- A i B - średnia roczna liczba szkód oszacowana na podstawie modelu M2,
- C i D - średnia roczna liczba szkód oszacowana na podstawie modelu M1.

Punkty „luźno rozrzucone" oznaczają rzeczywistą średnią roczną liczbę szkód wyznaczoną na podstawie zbioru uczącego.

....

Odp. b)

Należało wybrać model M2. Na taki wybór wskazuje mniejsza wartość kryterium AIC dla tego modelu (przy statystycznie istotnych wszystkich parametrach obydwu modeli) oraz zależność w kształcie litery „U” średniej rocznej liczby szkód wyznaczonej na podstawie zbioru uczącego od wieku (punkty „luźno rozrzucone”).

....

Odp. c)

Model M1

$$r_i^P = 3{,}021976876$$

Model M2

$$r_i^P = 2{,}894564838$$

Rozwiązanie:

Reszta Pearsona dla rozkładu Poissona:

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

Model M1

$\hat{\mu}_i = 0{,}090565115$, stąd $r_i^P = \frac{1-0{,}090565115}{\sqrt{0{,}090565115}} = 3{,}021976876$

Model M2

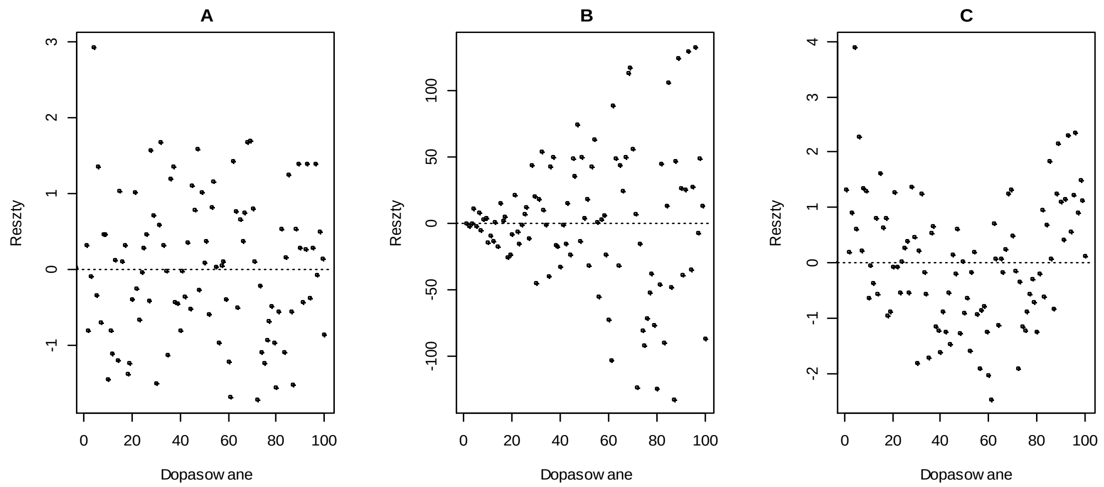
$\hat{\mu}_i = 0{,}097264522$, stąd $r_i^P = \frac{1-0{,}097264522}{\sqrt{0{,}097264522}} = 2{,}894564838$

Zadanie 3.

a) (***2p.***) Wymień cztery własności, które powinny spełniać reszty w liniowym modelu regresji.

b) (***1p.***) Na poniższym rysunku (Rys. 3.1) przedstawiono wykresy reszty vs wartości dopasowane (*Residuals vs Fitted*) dla trzech oszacowanych modeli A, B i C.
Co na podstawie tych wykresów można powiedzieć na temat reszt modeli A, B i C w kontekście własności omówionych w podpunkcie a)?

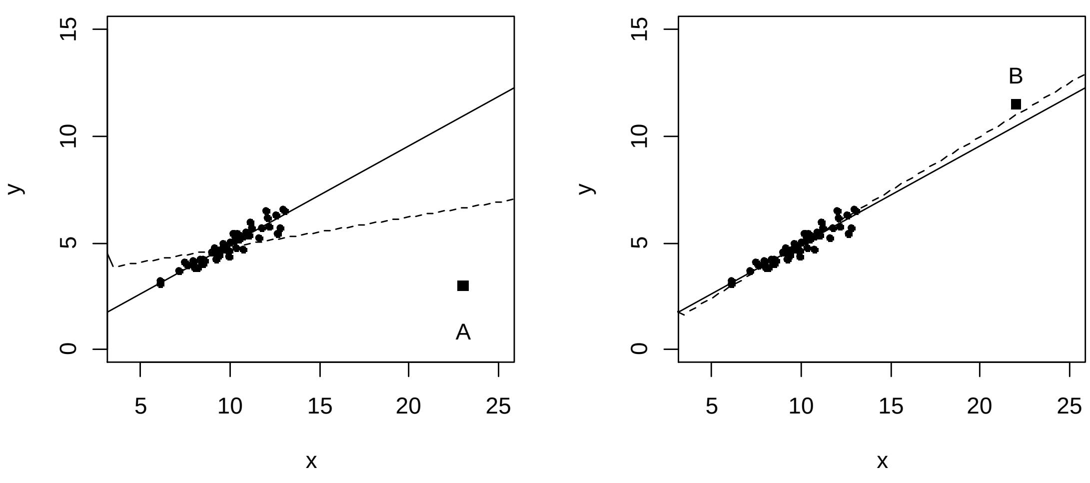
Rysunek 3.1



c) (***1p.***) Zdefiniuj obserwację wpływową (*influential observation*) i krótko omów w jaki sposób można ją zidentyfikować.

d) (***1p.***) Na wykresach zamieszczonych na poniższym rysunku (Rys. 3.2) linie ciągłe przedstawiają proste regresji wyznaczone bez uwzględnienia wyróżnionych punktów (tj. na lewym wykresie bez punktu A, a na prawym bez punktu B) a przerywane z ich uwzględnieniem. Czy obydwa punkty można uznać za wpływowe? Odpowiedź uzasadnij.

Rysunek 3.2



Odpowiedzi:

....

Odp. a)

Reszty poprawnie oszacowanego modelu liniowego (m.in.):

- powinny mieć średnią równą zero,
- powinny mieć rozkład normalny,
- powinny być jednorodne,
- nie powinny zależeć funkcyjnie od wartości,
- powinny być niezależne,
- nie powinny być skorelowane ze zmiennymi niezależnymi,

....

Odp. b)

Wykresy reszty vs wartości dopasowane wskazują, że

- reszty modelu A spełniają warunki poprawnie oszacowanego model liniowego,
- reszty modelu B nie są jednorodne,
- reszty modelu C zależą funkcyjnie od wartości dopasowanych.

....

Odp. c)

Obserwacja wpływowa to taka, której usunięcie ze zbioru danych spowodowałoby dużą zmianę dopasowania (wartości oszacowanych parametrów). Obserwacja taka może, ale nie musi, być wartością odstającą i może, ale nie musi, mieć dużą dźwignię, ale będzie miała co najmniej jedną z tych dwóch właściwości. W celu jej identyfikacji można wykorzystać miarę Cooka.

..

Odp. d)

Za wpływowy można uznać tylko punkt A. Usunięcie go ze zbioru danych powoduje znaczą poprawę dopasowania modelu.

Rozwiązanie:

Zadanie 4.

Wystąpienie oszustwa w zgłaszanych roszczeniach modelowano z wykorzystaniem uogólnionego modelu liniowego. Dla wartości progowej (punktu odcięcia) równej 0.25 otrzymano następującą macierz trafności prognoz (Tab. 4.1):

Tabela 4.1

Prognozy	Faktyczne	
	N ($y_i = 0$)	P ($y_i = 1$)
N ($y_i^P = 0$)	1200	160
P ($y_i^P = 1$)	60	80

a) (***1p.***) Wskaż jakie funkcje łączące (linki) mogą być wykorzystywane w uogólnionych modelach liniowych z binarną zmienną objaśnianą. Krótko wyjaśnij dlaczego takie funkcje są właściwe i podaj co najmniej trzy przykłady.
b) (***1p.***) Oblicz specyficzność (*specificity*) i czułość (*sensitivity*), wykorzystując podane w treści zadania dane.
c) (***2p.***) Narysuj krzywą ROC dla wartości progowej 0.25. Opisz osie oraz zaznacz i podaj współrzędne punktów, które odpowiadają wartościom progowym 0.25 oraz 0 i 1. Dodatkowo narysuj krzywe ROC dla
    i. modelu nie posiadającego żadnych zdolności predykcyjnych;
    ii. hipotetycznego „idealnego" modelu.
d) (***1p.***) Czy w tego typu modelach wysokość szkód może mieć wpływ na wybór wartości progowej (punktu odcięcia)? Odpowiedź krótko uzasadnij.

Odpowiedzi

....

Odp. a)

Za pomocą uogólnionych modelach liniowych z binarną zmienną objaśnianą modelujemy prawdopodobieństwo, że zmienna ta przyjmuje wartość jeden. Z kolei, predyktor liniowy może przyjmować dowolne wartości rzeczywiste. W związku z tym linkiem powinna być funkcja określona na przedziale $[0, 1]$ i przyjmująca wartości ze zbioru liczb rzeczywistych. Może to być np. logit, probit lub odwrotna dystrybuanty innego rozkładu określonego na zbiorze liczb rzeczywistych.

....

Odp. b)

$$\text{Specyficzność} = \frac{1200}{1260} = 0{,}9524$$

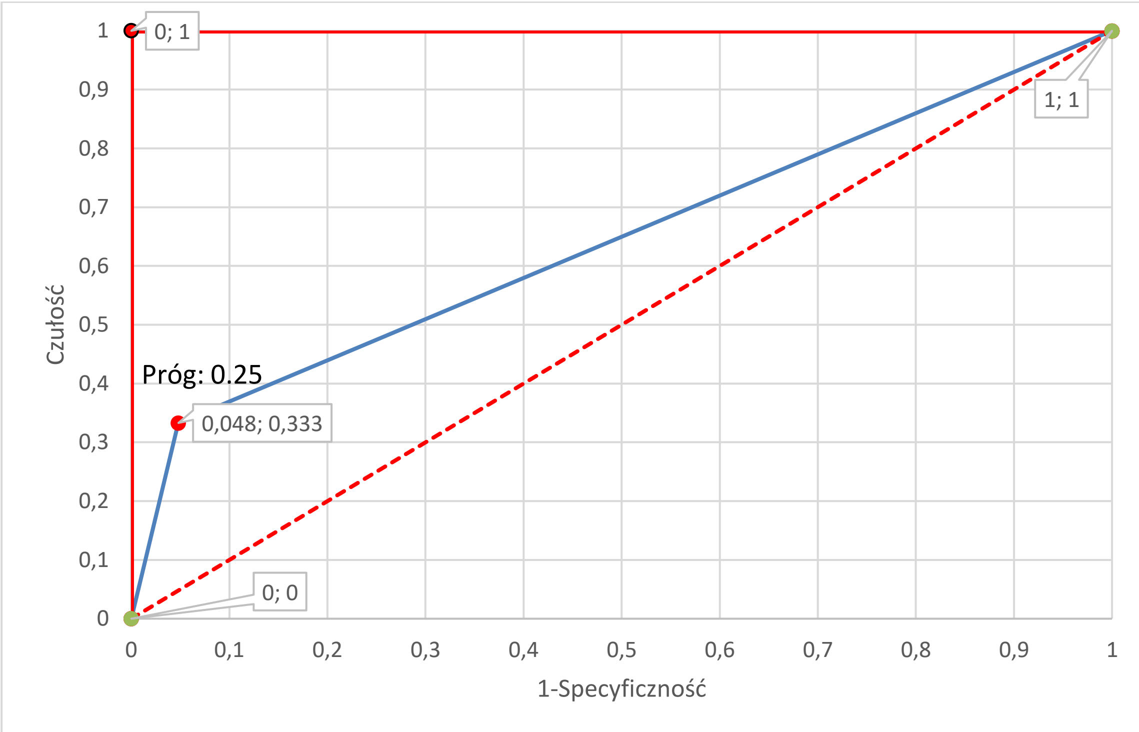
$$\text{Czułość} = \frac{80}{240} = 0{,}3333$$

....

Odp. c)

Krzywe ROC:

- niebieska dla wartości progowej 0.25
- czerwona przerywana dla modelu nie posiadającego żadnych zdolności predykcyjnych
- czerwona ciągła dla hipotetycznego „idealnego” modelu.



....

Odp. d)

Wysokość szkód może mieć wpływ na wybór wartości progowej. Z wartością progową związana jest liczba kontroli przypadków, dla których prognozowana jest możliwość oszustwa. O liczbie takich kontroli może decydować porównanie ich kosztów z kosztami niesłusznie wypłaconych odszkodowań. Np. w przypadku bardzo wysokich możliwych odszkodowań, koszty częstych kontroli mogą okazać się nisze od niesłusznie wypłaconych odszkodowań.

Rozwiązanie:

Zadanie 5.

Na podstawie tygodniowych logarytmicznych stóp zwrotu r_t dla spółki *Allianz* z okresu od 21-11-2003 do 06-05-2022 (liczba obserwacji T=964) oszacowano dwa modele M1 i M2:

– Model M1: ARMA(0,0)-GARCH(1,1) z rozkładem normalnym dla innowacji.
– Model M2: ARMA(0,0)-GARCH(1,1) ze skośnym rozkładem t-Studenta dla innowacji.

Uzyskano następujące wyniki:

Model **M1**

Optimal Parameters

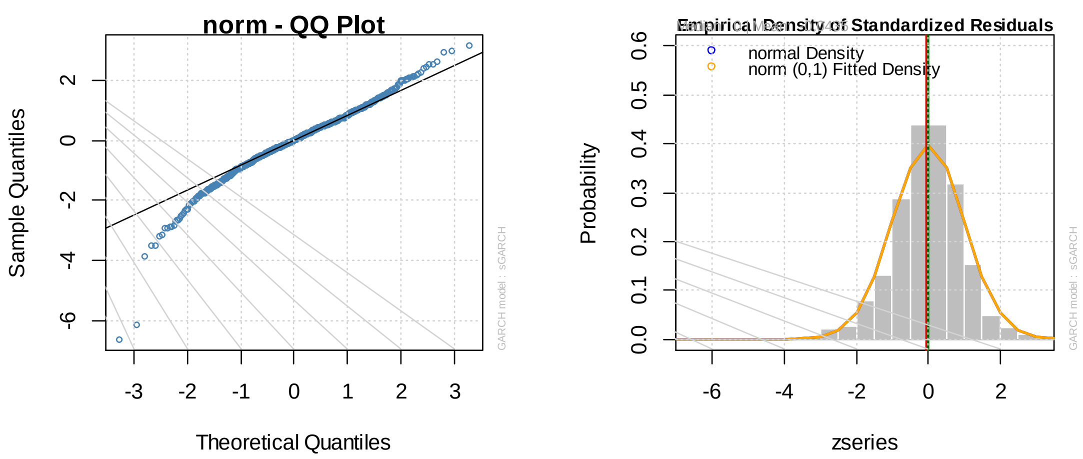
	Estimate	Std. Error	t value	Pr(>|t|)
mu	0.002223	0.001004	2.2135	0.026864
omega	0.000096	0.000024	3.9939	0.000065
alpha1	0.200018	0.036350	5.5026	0.000000
beta1	0.753919	0.036541	20.6324	0.000000

LogLikelihood : 1841.932

Information Criteria

Akaike	-3.8131
Bayes	-3.7929
Shibata	-3.8132
Hannan-Quinn	-3.8054

Rysunek 5.1



Wartości rzeczywiste i oszacowane warunkowe wariancje $\hat{\sigma}_t^2$ dla 3 ostatnich tygodni:

t	*962*	*963*	*964*
r_t	0.0004563085	-0.0137806499	-0.0556359297
$\hat{\sigma}_t^2$	0.0013652479	0.0011259237	0.0009960976

Model **M2**

Optimal Parameters

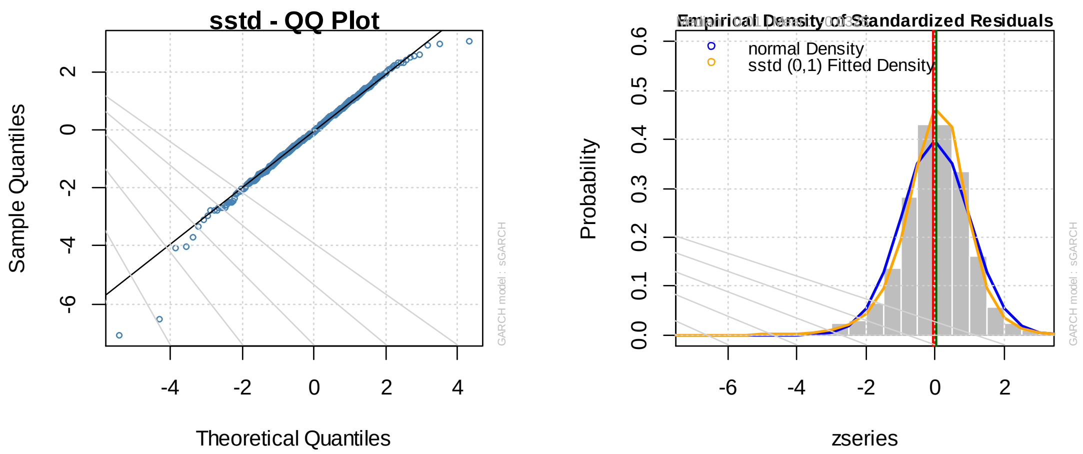
	Estimate	Std. Error	t value	Pr(>|t|)
mu	0.001836	0.000983	1.8671	0.061890
omega	0.000059	0.000021	2.8474	0.004407
alpha1	0.127545	0.030192	4.2245	0.000024
beta1	0.835771	0.032503	25.7138	0.000000
skew	0.880624	0.040562	21.7108	0.000000
shape	5.799026	1.020120	5.6847	0.000000

LogLikelihood : 1886.593

Information Criteria

Akaike	-3.9016
Bayes	-3.8713
Shibata	-3.9017
Hannan-Quinn	-3.8901

Rysunek 5.2



Wartości rzeczywiste i oszacowane warunkowe wariancje $\hat{\sigma}_t^2$ dla 3 ostatnich tygodni:

t	*962*	*963*	*964*
r_t	0.0004563085	-0.0137806499	-0.0556359297
$\hat{\sigma}_t^2$	0.001531988	0.001339137	0.001208819

a) (***2p.***) Krótko opisz klasę modeli GARCH(p, q).

b) (***2p.***) Wskaż, który z modeli M1 i M2 jest lepszy. Wybór uzasadnij, powołując się na podane wyniki ich oszacowań, w tym na wykresy zamieszczone na rysunkach 5.1 i 5.2.

c) (***1p.***) Wykorzystując wskazany model wyznacz prognozy warunkowej wariancji na okresy: $T+1,\ T+2$.

Odpowiedzi:

....

Odp. a)

Modele klasy GARCH wykorzystuje się do analizy szeregów czasowych. Stosuje się je głównie w analizie i prognozowaniu zmienności cen instrumentów finansowych. Za ich pomocą można opisać typowe własności finansowych szeregów czasowych.

....

Odp. b)

Należało wybrać model M2. Wskazują na to mniejsze wartości kryteriów informacyjnych, jak również lepsze dopasowanie skośnego rozkładu t-Studenta dla innowacji.

....

Odp. c)

$$\hat{\sigma}_{965}^2 = 0.001490078$$

$$\hat{\sigma}_{966}^2 = 0.001493918$$

Rozwiązanie:

Model ARMA(0,0)-GARCH(1,1) ma postać:

$$r_t = \mu + \varepsilon_t$$
$$\varepsilon_t = z_t \sigma_t$$
$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

Zatem wykorzystując model M2, otrzymujemy:

$$\hat{\varepsilon}_{964} = -0.0556359297 - 0.001836 = -0.05747166$$

$$\hat{\sigma}_{965}^2 = 0.000059 + 0.127545 \cdot -0.05747166^2 + 0.835771 \cdot 0.001208819$$
$$= 0.001490078$$

$$\hat{\sigma}_{966}^2 = 0.001493918$$

Zadanie 6.

a) (3***p***) Dane zgrupowano w następujący sposób:

x_i	n_i
$(c_0,\ \ c_1]$	n_1
$(c_1,\ \ c_2]$	n_2
$\vdots$	$\vdots$
$(c_{k-1},\ \ c_k]$	n_k
$(c_k,\ \ \infty)$	0

gdzie $c_0 = 0$. Przedstaw sposób wyznaczania dystrybuanty empirycznej dla danych zgrupowanych. Co to jest krzywa ogiwalna (*ogive*)? Jak się ją wyznacza?

b) (2***p***) Dane dotyczące wysokości 100 roszczeń (w tys. zł.) przedstawiono w następujący sposób:

x_i	n_i
(0, 1]	16
(1, 3]	22
(3, 5]	25
(5, 10]	18
(10, 25]	9
(25, 50]	6
(50, 100]	4
(100, ∞)	0

Wykorzystując krzywą ogiwalną wyznaczyć kwartyl 1 (Q_1) oraz decyl 9 (kwantyl rzędu 0.90).

Odpowiedzi:

....

Odp. a)

Strategia wyznaczania dystrybuanty dla danych zgrupowanych polega na obliczeniu dla końców przedziałów następujących jej wartości $F_n(c_j) = \frac{1}{n}\sum_{i=1}^{j} n_i$ (gdzie $n = \sum_{j=1}^{k} n_j$), a następnie połączeniu tych wartości w „rozsądny" sposób. Jeżeli połączymy punkty $(c_j, F_n(c_j))$ odcinakami otrzymujemy dystrybuantę empiryczną nazywaną krzywą ogiwalną. Wówczas:

$$F_n(x) = \frac{c_j - x}{c_j - c_{j-1}} F_n(c_{j-1}) + \frac{x - c_{j-1}}{c_j - c_{j-1}} F_n(c_j), \quad c_{j-1} \le x \le c_j.$$

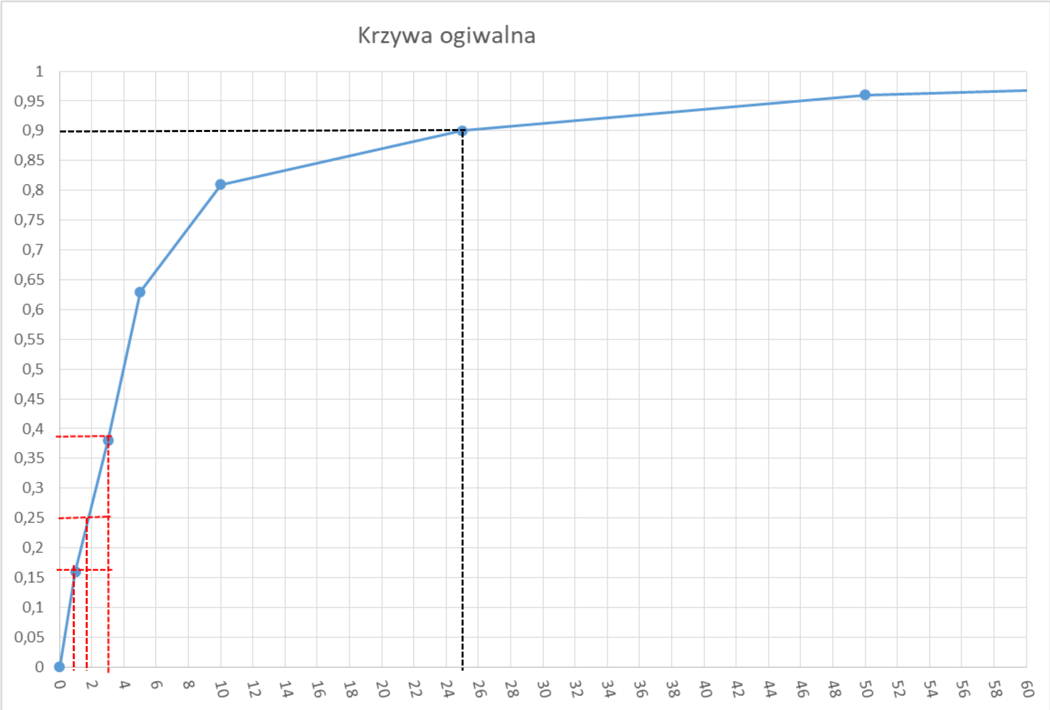
....

Odp. b)

$$q_{0.25} = Q_1 = 1.8181$$

$$q_{0.90} = 25$$

Rozwiązanie:



Kwartyl pierwszy:

$$\frac{0.38-0.16}{3-1}=\frac{0.25-0.16}{q_{0.90}-1}$$
$$q_{0.25}=Q_1=1.8181$$

$$q_{0.90}=25$$

Zadanie 7.

a) (***2p.***) Omów metodę symulacyjną wyznaczania rozkładu łącznych (zagregowanych) szkód w modelu ryzyka kolektywnego. Podaj sposób postępowania (algorytm).

b) (***3p.***) Wyznacz realizację z rozkładu łącznych (zagregowanych) szkód w modelu ryzyka kolektywnego przy założeniu, że liczba szkód ma rozkład Poissona z parametrem $\lambda = 2$ a wysokość pojedynczej szkody ma rozkład logarytmiczno-normalnym z parametrami $\mu = 1$ i $\sigma = 2$. Wykorzystaj po kolei niezbędną liczbę następujących wartości wylosowanych z rozkładu jednostajnego $U[0,1]$: 0.545, 0.773, 0.968, 0.739, 0.905, 0.960, 0.560

Odpowiedzi:

....

Odp. a)

W modelu ryzyka kolektywnego łączne szkody są modelowane za pomocą zmiennej $Z = \sum_{i=1}^{K} X_i$, gdzie: K – zmienna losowa opisując roczną liczbę szkód, X_i - zmienne losowe o takim samym rozkładzie opisujące wysokości pojedynczych szkód.
Algorytm:

1. Generujemy liczbę szkód *k*, przy założeniu określonej postaci rozkładu zmiennej *K*.
2. Generujemy wysokości k szkód: $x_1, \dots, x_k$, przy założeniu określonej postaci rozkładu zmiennej X_i.
3. Wyznaczamy sumę $z = x_1 + \cdots + x_k$, która jest jedną realizacją zmiennej *Z*.
4. Powtarzamy kroki (1)-(3) żądaną liczbę razy *n*.
5. Wyznaczamy dystrybuantę empiryczną $F_n(z)$ na podstawie pseudolosowej próby $z_1, \dots, z_n$.   .

....

Odp. b)

$$z = 122{,}5800$$

Rozwiązanie:

ad. b)

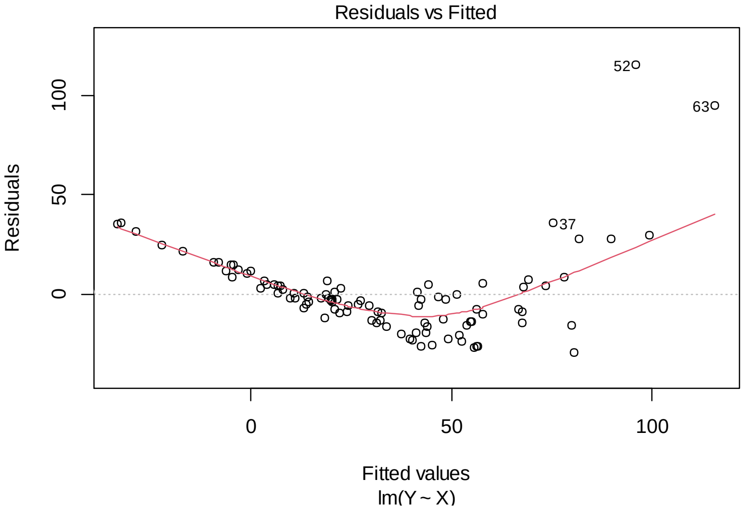
1. Losujemy liczbę z rozkładu Poissona z parametrem $\lambda = 2$. Korzystamy z metody odwracania dystrybuanty. Znajdujemy najmniejszą liczbę k (całkowitą nieujemną), która spełnia nierówność $F_K(k) \geq u$, gdzie u jest wartością wylosowaną z rozkładu jednostajnego $U[0,1]$. W rozważanym przypadku $u = 0.545$, $F_K(0) = 0.1353$, $F_K(1) = 0.4060$, $F_K(2) = 0.6767$. Tak więc wylosowana liczba szkód wynosi 2.
2. Losujemy dwie liczby z rozkładu logarytmiczno-normalnego z parametrami $\mu = 1$ i $\sigma = 2$. Korzystamy z metody odwracania dystrybuanty.
$F_{X_i}(x) = u$, czyli $\Phi^{-1}(u) = \frac{\ln(x) - \mu}{\sigma}$, stąd $F_{X_i}^{-1}(u) = \exp(\sigma\,\Phi^{-1}(u) + \mu)$
Stąd $x_1 = 12.1524$, $x_2 = 110.4276$.
3. Wyznaczamy realizację $z = 12.1524 + 110.4276 = 122{,}5800$.

Zadanie 8.

Analizując zależność między zmienną objaśnianą Y i zmienną objaśniającą X, w pierwszej kolejności oszacowano liniowy model regresji $y_i = \alpha + \beta x_i + \varepsilon_i$. Dla tego modelu:

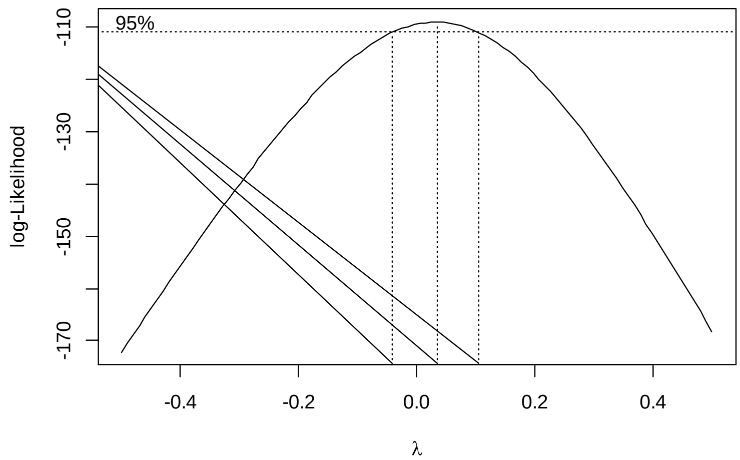
- Wyniki testu Breuscha-Pagana są następujące:
  studentized Breusch-Pagan test
  data: model.1
  BP = 12.419, df = 1, p-value = 0.000425
- Wykres reszty vs wartości dopasowane (*Residuals vs Fitted*) przedstawia poniższy rysunek (Rys. 8.1):

Rysunek 8.1



Na potrzeby analizy skonstruowano także przedstawiony na rysunku 8.2 wykres, w którym na osi odciętych są podane wartości parametru λ w transformacji Boxa-Coxa, a na osi rzędnych wartości funkcji wiarygodności modelu, w którym zmienna objaśniana X została przekształcona przez tą transformację (z parametrem λ).

Rysunek 8.2



a) (*1**p***) W jakim celu przeprowadza się transformacje zmiennej objaśnianej?
b) (*1**p***) Omów klasę transformacji Boxa-Coxa.

c) (*3**p***) Czy biorąc pod uwagę informacje podane w treści zadania, można twierdzić, że oszacowano adekwatny model? Jeżeli nie, to wskaż postać lepszego modelu. **Odpowiedzi uzasadnij powołując się na podane wykresy i wyniki testu.**

Odpowiedzi:

....

Odp. a)

Transformacje zmiennej zależnej przeprowadza się m.in. w celu:

- sprowadzenia zależności nieliniowych do postaci liniowej (wówczas często transformuje się również zmienne objaśniające),
- sprowadzenia jej rozkładu do rozkładu normalnego (zbliżonego do normalnego),
- stabilizacji wariancji.

....

Odp. b)

Klasę transformacji Boxa-Coxa określa się następująco:

$$y'_i = \begin{cases} \dfrac{y_i^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln(y_i), & \lambda = 0 \end{cases}, \qquad y_i > 0$$

Warunek $y_i > 0$ nie jest zbytnio krępujący, gdyż można wstępnie przesunąć zakres obserwowanych wielkości w obszar wartości dodatnich.

....

Odp. c)

Nie. Wskazują na to wyniki testu jednorodności wariancji Breuscha-Pagana i prawdopodobna funkcyjna zależność reszt od wartości dopasowanych (rys. 8.1).

Na podstawie wykresu przedstawionego na rys. 8.2, można przypuszczać, że otrzymamy lepszy model dla przekształconej zmiennej objaśnianej za pomocą transformacji Boxa-Coxa z parametrem λ bliskim zeru lub równym zeru (czyli dla $\ln(y_i)$).

Rozwiązanie:

Zadanie 9.

Dla pewnego portfela ubezpieczeń modelowano możliwość przedłużenia przez klienta polisy na kolejny rok. W tym celu, na podstawie danych zawartych w tabeli kontyngencji (Tab. 9.1), badano zależność między zmienną jakościową Y (przyjmującą dwie kategorie: *Tak* – klient przedłużył polisę, *Nie* – klient nie przedłużył polisy) oraz zmienną jakościową X określającą stan cywilny (przyjmującą trzy kategorie: M – w związku małżeńskim, S – singiel i P – inny).

Tabela 9.1

Y \ X	M	S	P
Tak	60	25	15
Nie	230	45	25

a) (***2p.***) Krótko scharakteryzuj metody badania siły zależności między zmiennymi jakościowymi mierzonymi na skali nominalnej. Jak można sprawdzić, czy związek między takimi zmiennymi jest statystycznie istotny?
b) (***2p.***) Wykorzystując odpowiedni test i dane z tab. 9.1, sprawdź czy między przedłużeniem przez klienta umowy a jego stanem cywilnym występuje istotny statystycznie związek. Przyjmij poziom istotności równy 0.05.
c) (***1p.***) Wykorzystując dane z tab. 9.1 oblicz współczynnik korelacji V Cramera.

Odpowiedzi:

....

Odp. a)

Badanie siły zależności między zmiennymi jakościowymi mierzonymi na skali nominalnej bazuje na wartości statystyki χ^2 wyznaczonej na podstawie tabeli kontyngencji. W oparciu o tą wartość skonstruowanych jest kilka mierników siły zależności, np. współczynnik φ Yule'a, współczynnik kontyngencji C Pearsona, współczynnik zbieżności T Czuprowa, współczynnik V Cramera.

Istotność sprawdzana jest testem niezależności chi-kwadrat.

....

Odp. b)

Występuje istotny statystycznie związek między przedłużeniem przez klienta umowy a stanem cywilnym.

....

Odp. c)

Współczynnik V Cramera jest równy 0,1620.

Rozwiązanie:

Ad. b)

Statystyka χ^2 wyraża się wzorem:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

gdzie:

n_{ij} – liczebności zaobserwowane,

$\hat{n}_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$ – liczebności teoretyczne (w tabeli na żółtym tle),

$n_{i\cdot}, n_{\cdot j}$ – liczebności brzegowe,

n – liczba obserwacji,

r, k – odpowiednio liczba wierszy i kolumn w tablicy kontyngencji.

Statystyka χ^2 ma rozkład Chi-Kwadrat o $(r-1)(k-1)$ stopniach swobody.

Obliczenia pomocnicze:

Y \ *X*	*M*	*S*	*P*	$n_{i\cdot}$
Tak	72.5 60	17.5 25	10.0 15	100
Nie	217.5 230	52.5 45	30.0 25	300
$n_{\cdot j}$	290	70	40	400

$$\chi^2 = 10.4926$$

Wartość krytyczna na poziomie istotności 0.05 wynosi 5.991. Czyli występuje istotny statystycznie związek między przedłużeniem przez klienta umowy a jego stanem cywilnym.

Ad. c)

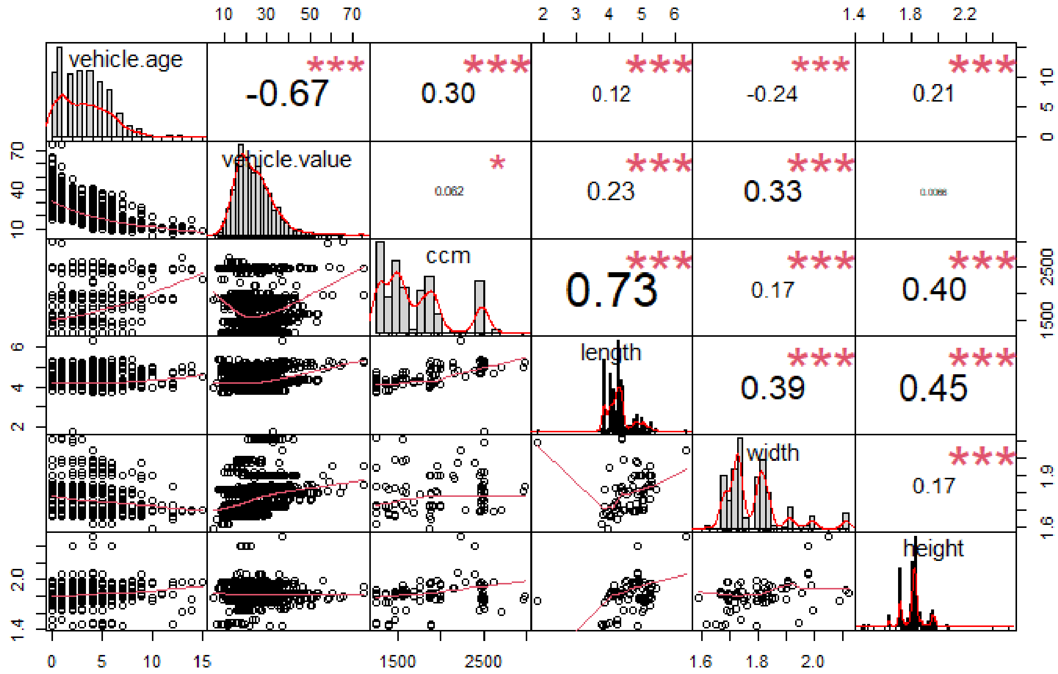
Współczynnik V Cramera wyraża się wzorem:

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

Stąd dla rozważanego przykładu $V = 0{,}1620$

Zadanie 10.
Ubezpieczone samochody opisuje sześć następujących zmiennych ilościowych: *vehicle.age* (wiek samochodu), *vehicle.value* (wartość samochodu), *ccm* (pojemność silnika), *length* (długość samochodu), *width* (szerokość samochodu), *height* (wysokość samochodu). Informacje o współczynnikach korelacji między nimi przedstawiono na rysunku 10.1. W celu redukcji liczby tych zmiennych zestandaryzowano je i przeprowadzono analizę składowych głównych. Uzyskano następujące wyniki dla 5-ciu wartości własnych macierzy korelacji: 1.8923127, 0.6867119, 0.6358262, 0.2488560, 0.2417949.

Rysunek 10.1



a) (*2p.*) Na czym polega analiza składowych głównych (*principal component analysis*)? Wskaż co najmniej trzy własności składowych głównych.
b) (*1p.*) Na podstawie wyników zaprezentowanych na Rys. 10.1 wypowiedz się na temat zasadności przeprowadzenie w analizowanym przypadku redukcji liczby zmiennych za pomocą składowych głównych.
c) (*2p.*) Jaką część wyjściowej wariancji reprezentują trzy pierwsze składowe główne?

Odpowiedzi:

....

Odp. a)
Analiza składowych głównych jest techniką eksploracji danych bez nadzoru. Polega na ortogonalnej transformacji układu badanych zmiennych X w zbiór nowych nieobserwowanych i niekorelowanych zmiennych Y. Nowe zmienne są liniowymi kombinacjami zmiennych obserwowanych. Po uporządkowaniu ich według malejącej wariancji, otrzymuje się główne składowe.

Głównych składowych można wyznaczyć tyle, ile było zmiennych pierwotnych. Jednak zazwyczaj kilka pierwszych wystarcza do wyjaśnienia większości wariancji układu zmiennych.
Własności składowych głównych (m.in.):

- są liniową kombinacją obserwowanych zmiennych,
- są ortogonalne względem siebie,
- kolejne składowe wyjaśniają malejąca ilość łącznej wariancji zmiennych,
- suma wariancji składowych jest równa sumie wariancji zmiennych pierwotnych.

....

Odp. b)
Redukcje liczby zmiennych przeprowadza się, gdy wszystkie lub kilka z nich jest ze sobą istotnie skorelowanych. W analizowanym przypadku mamy do czynienia z taką sytuacją, więc przeprowadzenie redukcji liczby zmiennych za pomocą składowych głównych można uznać za zasadne.

....

Odp. c)
Trzy pierwsze składowe główne reprezentują 81.22% wariancji.

Rozwiązanie:

Ponieważ mamy 6 zestandaryzowanych zmiennych, więc łączna wariancja jest równa 6.
Wariancja (wartość własna) brakującej składowej wynosi:
$6 - (1.8923127 + 0.6867119 + 0.6358262 + 0.2488560 + 0.2417949) = 2.2944983$
Jest ona największa, więc brakująca składowa jest pierwszą składową główną.
Zatem trzy pierwsze składowe główne tłumaczą $\frac{2.2944983+1.8923127+0.6867119}{6} 100\% = 81.22\%$ wariancji.

Sesja egzaminacyjna w dniu 9 czerwca 2022 r.

Modelowanie

Arkusz ocen

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