Analiza doboru modelu regresji dla rozkładu Poissona na przykładzie analizy ryzyka awarii[1]

Dodatek do Rozdziału 1 skryptu:

„Metoda największej wiarygodności i informacja Fisher'a w fizyce i ekonofizyce"

Jacek Syska
Instytut Fizyki, Uniwersytet Śląski

[1] (wersja pierwsza)

Spis treści

Wstęp

Dodatek ten jest uzupełnieniem do Rozdziału pierwszego skryptu [1] „Metoda największej wiarygodności i informacja Fisher'a w fizyce i ekonofizyce". Celem dodatku jest krótkie, praktyczne wyjaśnienie działania metody największej wiarygodności (MNW) oparte o przykład analizy doboru modelu dla regresji Poissona, z wykorzystaniem możliwości procedur zawartych w pakiecie SAS (system analiz statystycznych). Podstawy teoretyczne MNW oraz aparatu matematycznego związanego z zastosowaniem informacji Fishera może czytelnik znaleźć między innymi w pozycji [1].

MNW jest ogólną statystyczną metodą otrzymywania estymatorów parametrów populacyjnych modelu statystycznego. Estymatory MNW mają dla dużej próbki optymalne właściwości statystyczne [2]. Dla małej próbki skorzystanie z pełni praktycznych zalet MNW możliwe jest dopiero po odwołaniu się do formalizmu geometrii różniczkowej na przestrzeni statystycznej modeli statystycznych [3, 1].

Zaletą MNW w estymacji parametrów jest to, że można ją zastosować w rozmaitych sytuacjach. Jej ważną cechą jest to, że ogólne zasady i procedury mogą być używane do przeprowadzania wnioskowania statystycznego dla modeli regresji ze zmienną objaśnianą o dowolnym rozkładzie. Stąd to samo wnioskowanie statystyczne MNW może być (z dokładnością do różnic modelowych) zastosowane w analizie regresji np. klasycznego modelu normalnego, jak i w analizie regresji Poissona.

Gdy model wielorakiej regresji liniowej jest dopasowany do danych empirycznych zmiennej objaśnianej posiadającej rozkład normalny, wtedy estymatory współczynników regresji metody najmniejszych kwadratów (MNK) są identyczne jak estymatory otrzymane w MNW [1]. Estymacja MNW parametrów modelu umożliwia również analizę modeli nieliniowych, takich jak np. model regresji logistycznej [4] oraz rozważany w niniejszym Dodatku model regresji Poissona. Zrozumienie działania MNW w estymacji parametrów i umiejętność dokonywania wyboru modelu w oparciu o odpowiednie testy statystyczne jest niezbędną umiejętnością współczesnych analiz statystycznych w wielu dziedzinach nauk empirycznych.

Analiza regresji Poissona jest stosowana w modelowaniu zależności pomiędzy zmiennymi w przypadku, gdy zależna zmienna losowa (nazywana też zmienną opisywaną lub odpowiedzią) przyjmuje z natury tej zmiennej realizacje w postaci zbioru dyskretnych danych. Na przykład zmienna objaśniana może być liczbą zliczeń przypadków interesującego nas zdarzenia, np. liczbą przypadków awarii, które pojawiają się w ustalonym czasie badania.

Dla typowego modelu regresji Poissona naturalną miarą estymowanego defektu jest ryzyko względne, związane z określonym, interesującym nas czynnikiem.

Celem Dodatku jest wyjaśnienie jak postulować i badać postać modelu regresji Poissona oraz jak wykorzystywać kluczowe cechy modelu do estymacji parametru ryzyka względnego, kontrastującego porównywane zbiorowości ze względu na warianty czynników ryzyka. W Dodatku wykorzystamy wprowadzone w skrypcie [1] pojęcia statystyki ilorazu wiarygodności oraz dewiancji, stosując je do analizy selekcji modelu właściwego dla przykładowych danych (których realizacja jest możliwa), co do których uznamy [5], że pochodzą z rozkładu Poissona. W Dodatku przedstawiony zostanie typowy model regresji Poissona, który wyraża w postaci logarytmicznej tempo porażki (np. awarii) jako liniowej funkcji zbioru czynników. Metoda regresji Poissona, może być również zastosowana w bardziej skomplikowanych nieliniowych modelach. Zainteresowanego czytelnika odsyłamy do [4].

MNW na przykładzie analizy modeli regresji Poissona

Poniżej przedstawimy działanie MNW w estymacji parametrów modelu regresji Poissona [1] *oraz pokażemy jak połączyć wnioskowanie statystyczne z tworzeniem odpowiedniej procedury pakietu SAS.*

Analizę regresji Poissona stosuje się w modelowaniu zachowania się zmiennej objaśnianej przyjmującej, z natury tej zmiennej, dyskretne realizacje widoczne w danych i powstałe np. ze zliczeń modyfikowanych zmiennymi objaśniającymi (nazywane czynnikiami). Po pierwsze, wyjaśnimy jak konstruować postać modelu regresji Poissona dla tzw. ryzyka względnego i jak MNW dokonuje estymacji parametrów modelu. Wnioskowanie związane z weryfikacją hipotez o braku dopasowania w modelu niższym przedstawimy w drugiej kolejności.

D1.1 Rozkład Poissona

Rozkład Poissona jest często używany do modelowania zjawisk pojawiania się rzadkich zdarzeń, takich jak nowych przypadków awarii w pewnej populacji w pewnym okresie czasu albo zajścia określonej liczby wypadków samochodowych w pewnym określonym miejscu w ciągu roku.

Przyjmijmy więc, że zmienna objaśniana jest zmienną losową Y przyjmującą wartości y zgodnie z rozkładem Poissona:

$$p(y \mid \mu) = \frac{\mu^y e^{-\mu}}{y!} \quad , \qquad y = 0,1,...,\infty \quad , \tag{1}$$

gdzie μ jest parametrem rozkładu. Zmienna losowa Poissona Y może przyjąć dowolną, nieujemną wartość całkowitą y. Na przykład, zgodnie z (1) prawdopodobieństwo, że Y przyjmuje wartość $y = 7$ wynosi:

$$pr(y = 7 \mid \mu) = \frac{\mu^7 e^{-\mu}}{7!} = \frac{\mu^7 e^{-\mu}}{5040} \ .$$

Widać, że prawdopodobieństwo to zmienia się jako funkcja wartości parametru μ. W analizie MNW koncentrujemy się na badaniu zależności rozkładu prawdopodobieństwa zmiennej objaśnianej, od parametrów tego rozkładu.

Rozkład Poissona posiada interesującą statystycznie własność [2, 1]:

$$E(Y) = \sigma^2(Y) = \mu \ . \tag{2}$$

D1.2 Przykład danych dla regresji Poissona

Aby zilustrować działanie MNW w analizie regresji Poissona rozważmy dane przedstawiające awarię urządzenia określonego typu (pomijając awarię niszczącą całkowicie urządzenie). Tego typu analiza została zastosowana ze sporym sukcesem w badaniach medycznych [4].

Poniższa Tabela 1 przedstawia dwie *przykładowe* próbki pobrane z populacji silników serwisowanych samochodów pewnej firmy (nazwijmy ją „Auto") i jej modelu typu „Model", które uległy niedestrukcyjnej awarii, tzn. takiej, po której silnik można jeszcze naprawić nie zmniejszając tym samym wielkości populacji, z których dokonujemy losowania.

Próbki powstały na skutek losowania pewnej liczby aglomeracji miejskich i takiej samej liczby aglomeracji wiejskich na całym obszarze ziemi, na którym firma „Auto" ma swój serwis. Próbki w obszarach Miejskim i Wiejskim zostały uporządkowane wg wariantów wieku (miesięcy używania).

W przykładzie zmienna zależna Y jest zmienną zliczeń przypadków awarii silnika. Generalne populacje dwóch obszarów używania samochodów zakwalikowano do ośmiu wariantów wiekowych. Stąd zmienną Y indeksujemy podwójnym indeksem grupowym, tzn. Y_{ij} oznacza liczbę zliczeń dla i–tego wariantu wiekowego i j–tego obszaru, gdzie i zmienia się od 1 do 8, natomiast $j = 0$ dla obszaru „Miasta" oraz $j = 1$ dla obszaru „Wsie". Oznaczmy przez ℓ_{ij} rozmiar podpopulacji dla i–tego wariantu wieku samochodu i j–tego obszaru.

Celem analizy jest ustalenie, czy ryzyko awarii silnika samochodu, przy dopasowaniu ze względu na wiek, jest wyższe w pierwszym badanym obszarze czy w drugim.

D1.2.1 Rola kowarianta

„Wiek" jest wspólną *zmienną poboczną* dla obu rozważanych populacji, tzw. *kowariantem* zmiennej „obszar". Należy wprowadzić go do analizy bądź w *członach interakcji* ze zmienną „obszar" lub jako *zaburzeniem* wpływu głównego, którym jest zmienna „obszar" [4]. Wprowadzenie „wieku" do analizy oznacza, że zmienna ta jest pod kontrolą oraz, że oszacowany parametr, którym w naszym przykładzie okaże się być ryzyko względne, jest estymowany w sytuacji dopasowania zmiennych i estymatorów parametrów modelu ze względu na zmienną „wiek" samochodu. Pominięcie kowarianta oznaczałoby wyznaczanie *surowych* estymatorów *parametrów*.

D1.3 Pojęcie ryzyka

Termin ryzyko w rozważanym przykładzie odnosi się do rozwijającego się z czasem prawdopodobieństwa zajścia wady silnika. **Przez** r_{ij} **będziemy oznaczać rzeczywiste populacyjne ryzyko w grupie** (i, j).

D1.3.1 Analogia ryzyka awarii i prawdopodobieństwa zajścia porażki na jednostkę czasu. Estymowane tempo defektu

Rozważmy rozkład dwumianowy z parametrem prawdopodobieństwa p oraz liczby losowań m. Związek określający oczekiwaną liczbę sukcesów w m losowaniach Bernoulliego $\mu = m\,p$, można zapisać następująco:

$$\mu = (m \cdot \Delta t)\frac{p}{\Delta t} \quad ,$$

skąd widać, że jeśli Δt jest czasem prowadzonego badania, wtedy $l \equiv m \cdot \Delta t$ jest zakumulowaną w tym czasie liczbą „samochodo–miesięcy", a $r \equiv \frac{p}{\Delta t}$ jest tzw. **intensywnością,** *czyli prawdopodobieństwem zajścia zdarzenia na jednostkę czasu, nazywanym ryzykiem*.

Pojęcie ryzyka: Ze względu na to, że μ jest liczebnością, związek $\mu = (m \cdot \Delta t)\frac{p}{\Delta t}$ ma postać analogiczną (aczkolwiek jedynie analogiczną) do stosowanej w analizie regresji Poissona postaci funkcji regresji $\mu_{ij} = \ell_{ij} r_{ij}$ [4, 1] dla wartości oczekiwanej μ_{ij} liczby zliczeń zdarzeń awarii w grupie (i, j), gdzie ℓ_{ij}, który jest odpowiednikiem $(m \cdot \Delta t)$, jest parametrem określającym liczbę wszystkich wyników zakumulowanych w czasie badania.

R**yzyko** w grupie (i, j) jest zdefiniowane jako:

$$r_{ij} = \frac{\mu_{ij}}{\ell_{ij}} \quad . \tag{3}$$

Jest ono *analogiem* **intensywności** $r = \frac{\mu}{(m \cdot \Delta t)}$.

Estymowane ryzyko nazywane **tempem defektu** rozumianego jako porażka, jest ogólnie definiowane jako:

$$\hat{r}_{ij} = \frac{Y_{ij}}{\ell_{ij}} \quad , \tag{4}$$

gdzie Y_{ij} jest ilością zaobserwowanych zliczeń defektów silnika dla podgrupy (i, j), a ℓ_{ij} oznacza zakumulowaną (tzn. sumaryczną) długość wolnego od defektu czasu dla wszystkich samochodów w tej podgrupie. Zatem $\hat{r}_{ij}$ mierzy liczbę defektów w stosunku do całkowitej zakumulowanej liczby wszystkich samochodów poddanych serwisowaniu w danej podgrupie na ustaloną jednostkę czasu (np. roku). Zwróćmy uwagę, że występująca w liczniku (4) zmienna Y_{ij} nie jest w ogólności estymatorem MNW parametru μ_{ij} dla modelu regresji Poissona, chociaż jest tak dla modelu podstawowego.

D1.3.3 Ryzyko względne

Stosunek:

$$R_{Wi} = \frac{r_{i1}}{r_{i0}} \tag{5}$$

jest parametrem nazywamy *ryzykiem względnym* lub *ilorazem ryzyk*, który w tym przypadku jest stosunkiem r_{i1} dla populacji „Wiejskiej" w i–tym wariancie wiekowym do ryzyka r_{i0} dla populacji „Miejskiej", również w i–tym wariancie wiekowym.

Jeżeli $R_{Wi} = 1$, to ryzyka populacyjne są takie same w obu i–tych wariantach wiekowych, jeżeli $R_{Wi} > 1$, to ryzyko dla Wsi jest wyższe niż dla Miast w danym wariancie wieku samochodu.

Alternatywne nazwy ryzyka względnego.

Innymi używanymi określeniami ryzyka względnego $R_{Wi} = r_{i1} / r_{i0}$ są: stosunek temp, stosunek intensywności (IDR), iloraz zapadalności, stosunek częstości, iloraz prawdopodobieństw lub po prostu, stosunek ryzyk.

D1.4 Uwaga o ogólnym indeksowaniu podgrup populacji

W ogólnych rozważaniach, każda wartość indeksu grupowego j=1,2,...,N, wskazuje i-tą (generalną) populację, w której (nielosowe) czynniki X_i, i=1,2,...,p, przyjmują ustalone

wartości, im właściwe. W ten sposób liczba wszystkich (pod)populacji wskazanych indeksem i oraz wartościami zmiennych X_i, i=1,2,...,p wynosi $N \times z_1 \times z_2 \times ... \times z_p$, gdzie z_i jest liczbą dyskretnych wartości, które przyjmuje zmienna X_i. W każdej z tych podpopulacji zmienną losową Y oznaczamy jako $Y_{l_1,...,l_p,j}$, gdzie l_i = 1,...,z_i dla i=1,2,...,p, a jej zmierzoną wartość jako $y_{l_1,...,l_p,j}$. Zbiór wszystkich $Y_{l_1,...,l_p,j}$ tworzy próbę oznaczaną tak jak poprzednio przez $\tilde{Y}$.

D1.5 Dane dla przykładu

Tabela 1 danych dla przykładu: Porównanie wystąpienia awarii silnika samochodów „Model" firmy „Auto" użytkowanych przez mieszkańców obszarów Miejskich oraz Wiejskich na całym obszarze dostępnym przez serwis tej firmy. Liczebności występujące w tabeli w są sumarycznymi liczebnościami dla próbki powstałej z wszystkich wylosowanych aglomeracji Miejskich (lub Wiejskich).

Wiek grupy samochodów (w miesiącach)	Obszary Miejskie		Obszary Wiejskie		Estymowany wskaźnik ryzyka, gdzie obszar Miast jest grupą referencyjną
	Ilość przypad-ków	Rozmiar próbek serwisowanych samochodów	Ilość przypadków	Rozmiar próbek serwisowanych samochodów	
0 – 12	1	172675	4	181343	3,81
13 – 24	16	123065	38	146207	2,00
25 – 36	30	96216	119	121374	3,14
37 – 48	71	92051	221	111353	2,57
49 – 60	102	72159	259	83004	2,21
61 – 72	130	54722	310	55932	2,33
73 – 84	133	32185	226	29007	1,89
85 +	40	8328	65	7538	1,80

Uwaga: Dla danych w Tabeli 1 dotyczących jednego wariantu wielu badań serwisowych w populacji „Miejskiej" (j=0) lub „Wiejskiej" (j=1), jedna liczba w kolumnach 3 lub 5 podająca rozmiar próbki, jest rozumiana jako liczba samochodo–miesięcy w czasie prowadzonego badania dla określonego j-tego obszaru i i-tego wariantu wieku podgrupy (i, j), gdzie i =1,2,...,8.

D1.5.1 Cel badań

W ostatniej kolumnie Tabeli 1 podano *estymowane* z pobranych próbek ryzyka względne, w każdym z wariantów wiekowych. W każdym wariancie wieku ryzyka wyniosły więcej niż 1, co jasno sugeruje, że obszar Wiejski ma wyższy ogólny wskaźnik awaryjności niż Miejski.

Analiza regresji Poissona ma ustalić, czy powyższy „na oko" widoczny wzorzec danych w Tabeli 1 jest statystycznie istotny oraz otrzymać estymator ogólnego ryzyka względnego, który byłby dopasowany ze względu na wiek samochodu (tzn. wiek samochodu jest zmienną pod kontrolą).

D1.5.2 Uzasadnienie zastosowania rozkładu Poissona w analizie.

Fakt, że rozkład Poissona jest użyteczny dla modelowania pewnych typów zliczeń zdarzeń dla danych serwisowych, jest oparty na tym, że rozkład Poissona jest przybliżeniem rozkładu dwumianowego *B* [7]. Ściśle rzecz biorąc, rozkład dwumianowy $B(m, p)$ przechodzi w rozkład Poissona $Poisson(\mu = m\,p)$ zmiennej Y tylko granicznie wtedy, gdy przy rozmiarze populacji *m* dążącym do nieskończoności i dwumianowym parametrze prawdopodobieństwa *p* bardzo małym, wartości oczekiwana liczby zdarzeń $\mu = E(Y) = m\,p$ pozostaje ustalona na wartości oczekiwanej rozkładu dwumianowego [7]. W granicy tej oczekiwana dwumianowa liczba zliczeń (wartość oczekiwana μ) jest względnie mała w porównaniu z rozmiarem populacji, a rozkład Poissona daje dobre przybliżenie rozkładu dwumianowego dla rzadkich przypadków awarii.

Dlatego zastosowanie modelu Poissona jest sugerowane, gdy otrzymujemy dużą liczbę wszystkich wyników dla próbki pobranej z populacji, w której bada się rozwój awaryjności, np. rozwój rzadkiej awarii silnika, tak że wielkość zakumulowanego (samochodo-)czasu jest duża, a jednocześnie tempo r_{ij} pojawiania się interesujących nas zdarzeń jest małe.

Dane w Tabeli 1 zadowalająco dobrze spełniają to założenie, gdyż w każdej kategorii wiekowej występuje stosunkowo mały udział względny przypadków awarii w porównaniu do rozmiaru odpowiedniej podpopulacji. Jednak pełna analiza powinna obejmować test nieparametrycznej hipotezy o typie rozkładu, z którego generowane są dane [5]. Sprawdzenie tego faktu pozostawiamy czytelnikowi jako ćwiczenie.

D1.5.3 Przykład fizycznego odpowiednika danych w przykładzie.

Pojęcie „serwisowego” ryzyka względnego nie jest niepodobne do żadnej wielkości pojawiającej się np. w modelach fizycznych. Jej fizycznym odpowiednikiem jest iloraz przekrojów czynnych stosowany do opisu zajścia badanego procesu, który jest typem kontrastu różnych możliwych kanałów zachodzącej reakcji. W przypadku, gdy zmienna objaśniana ma pewien rozkład z wartością oczekiwaną zmieniającą się w zależności od wariantów zmiennej głównej oraz zmiennych pobocznych (kowariantów), wtedy w przypadku braku interakcji zmiennej głównej ze wspomnianymi kowariantami, zastosowanie stosunku temp może być przyczyną „zniknięcia” wpływu tych drugich na wartość ilorazu. W przypadku braku interakcji sytuacja ta byłaby więc podobna do omówionej w D1.8.

D1.6 Postać funkcji regresji pierwszego rodzaju

Zbudowanie modelu regresji Poissona dla powyższej sytuacji oznacza opisanie oczekiwanej liczby przypadków awarii silnika, $E(Y_{ij})$, poprzez wprowadzone do modelu zmienne objaśniające (tzw. czynniki). Zauważmy, że liczba zliczeń Y_{ij} jest teoretycznie zmienną losową dwumianową z wartością oczekiwaną równą $\mu_{ij} = \ell_{ij} r_{ij}$. Równanie to wyraża treść funkcji regresji pierwszego rodzaju, tzn. postulowaną jej postać w całej generalnej populacji.

D1.6.1 Indeksowanie grup w przykładzie

Analizę dla regresji Poissona, omówimy na przykładzie danych z Tabeli 1 opisujących liczbę niedestrukcyjnych awarii silnika dla samochodów sklasyfikowanych wg wariantów wiekowych w Miastach i Wsiach.
Zgodnie z już wcześniej wprowadzonymi oznaczeniami, ze względu dwie populacje „Miast" i „Wsi", liczba generalnych populacji N=2 skąd, ze względu na poniżej wprowadzone kodowanie zmiennych ukrytych, przyjmujemy j=0,1. Natomiast ze względu na występowanie jednego czynnika „wieku" samochodu X=X_1 , indeks i=p=1. W danych z Tabeli 1, (kategoryzujący) czynnik X przyjmuje z = 8 wartości. Stąd liczba wszystkich podpopulacji wynosi $N \times z = 2 \times 8 = 16$, a każdą z podpopulacji (podgrup) wskazuje para indeksów grupowych (i , j). Zmienne losowe Y oznaczamy jako Y_{ij}, gdzie i = 1,...,8 , a j=0,1. Indeksowanie dla populacji i podpopulacji przenosi się automatycznie na indeksowanie pobranych z tych populacji próbek.

D1.7 Równanie regresji Poissona ze zmiennymi ukrytymi

W rozważanym przykładzie występują dwa czynniki, czynnik wpływu głównego, którym jest „obszar" serwisowania oraz czynnik poboczny „wieku" samochodu. Ponieważ „wiek" będzie klasyfikowany w ośmiu kategoriach, użyjemy do ich wskazania (indeksowania) siedmiu zmiennych ukrytych [4]. Zmienna „obszar", która zawiera dwa warianty, wymaga tylko jednej zmiennej kierunkowej.

Ogólna postać modelu regresji, czyli funkcji opisującej zmianę wartości oczekiwanej liczby awarii (silnika) wraz ze zmianą grupy (i, j), może być zapisana następująco [4]:

$$E(Y_{ij}) = \mu_{ij} = \ell_{ij} r_{ij} \quad , \qquad i = 1, 2, \ldots, 8 ; \quad j = 0, 1 . \tag{6}$$

Wspomniane **zmienne ukryte** (kierunkowe) U_k oraz M wskazującą w następujący sposób [4] odpowiednio wariant „wieku" oraz „obszaru":

$$U_k = \begin{cases} 1 & \text{jeśli} \quad k = i , \qquad \text{gdzie} \quad i = 1, 2, \ldots, 7 \\ 0 & \text{w przeciwnym wypadku} \end{cases} \tag{7}$$

$$M = \begin{cases} 1 & \text{jeśli} \quad j = 1 \qquad \text{(Wsie)} \\ 0 & \text{jeśli} \quad j = 0 \qquad \text{(Miasta)} \quad . \end{cases} \tag{8}$$

Podstawowa dla wielu analiz regresji Poissona, logarytmiczna postać funkcji ryzyka [6] występująca w (6) i korzystająca z kodowania (7) oraz (8) ma w przypadku **bez interakcji** następującą postać:

Model 1: $$\ln r_{ij} = \alpha + \sum_{k=1}^{7} \alpha_k U_k + \beta M \quad . \tag{9}$$

Korzystając z kodowania (7) i (8), możemy w powyższym „Modelu 1" ryzyka, wyrazić r_{ij} poprzez parametry α_i i β w następujący sposób:

$$\ln r_{i0} = \alpha + \alpha_i \qquad \text{oraz} \qquad \ln r_{i1} = \alpha + \alpha_i + \beta \ , \qquad i = 1, 2, \ldots, 7, \tag{10}$$

oraz

$$\ln r_{80} = \alpha \qquad \text{oraz} \qquad \ln r_{81} = \alpha + \beta \quad , \qquad \text{dla } i = 8 , \tag{11}$$

co wynikało z tego, że $U_k = 1$ dla $k = i = 1, 2, \ldots, 7$ oraz $U_k = 0$ dla $i = 8$.

Powyższy przykład modelowania jest wykorzystywany w estymacji ryzyka rozwijania się uszkodzenia (silnika samochodu) z wiekiem. Bardziej ogólne i popularne zastosowania regresji Poissona dotyczą modelowania tempa defektów, czyli tzw. *intensywności* procesu, dla różnych interesujących nas podgrup.

Wniosek: Ze związków (10) i (11) widzimy, że w traktowanych osobno obszarach „Miejskim" i „Wiejskim" ryzyko (tempo awarii) r_{ij} zmienia się z wariantem „wieku", co z powodu niezerowych oszacowań współczynników α_i będzie widoczne w poniższych raportach SAS.

Uwaga o alternatywnym kodowaniu:
Alternatywnie model może być zdefiniowany poprzez użycie ośmiu zmiennych kierunkowych dla wieku i jednej zmiennej kierunkowej dla obszaru [4]. Gdyby zastosowano osiem zmiennych kierunkowych dla wieku, użycie wyrazu wolnego byłoby błędem.

D1.8 Estymator ogólnego ryzyka względnego w modelu bez interakcji

Poniżej wyprowadzimy ważny wniosek dotyczący ryzyka względnego w modelu bez interakcji czynnika „obszar" z czynnikiem pobocznym „wieku".

Korzystając z (10) i (11) otrzymujemy:

$$\ln r_{i1} - \ln r_{i0} = (\alpha + \alpha_i + \beta - \alpha - \alpha_i) = \beta \quad , \qquad i = 1, 2, ..., 7 \qquad (12)$$

oraz

$$\ln r_{81} - \ln r_{80} = (\alpha + \beta - \alpha) = \beta \ , \qquad i = 8 \ . \qquad (13)$$

Korzystając z (12) oraz (13) widzimy, że **ryzyko względne** (4) dla modelu (9) nie zawierającego interakcji jest równe:

$$R_{Wi} = \frac{r_{i1}}{r_{i0}} = \exp\left[\ln\left(\frac{r_{i1}}{r_{i0}}\right)\right] = \exp[\ln r_{i1} - \ln r_{i0}] = \exp[\beta] = e^{\beta} \quad , \qquad i = 1, 2, ..., 8. \qquad (14)$$

Powyższy model pozwala na estymację wskaźnika ryzyka względnego dla każdej kategorii wiekowej. Czynimy to metodą MNW estymując współczynnik kierunkowy β przy zmiennej *M* i w ten sposób dopasowując model, a następnie licząc eksponentę tego estymatora. Ponieważ estymowany wskaźnik ryzyka względnego e^{β} jest niezależny od *i* (tzn. od kategorii wiekowej), zatem możemy interpretować $e^{\hat{\beta}}$ jako *estymator ogólnego ryzyka względnego*, dopasowanego do wieku, gdzie $\hat{\beta}$ jest estymatorem MNW parametru β.

Wniosek *o postaci ryzyka względnego w modelu bez interakcji*: **Dla modelu** (9) **bez interakcji zmiennych „obszar" i „wiek"** (oznaczonego jako Model 1), **ryzyko względne nie zależy od wariantu wiekowego**, tzn. wpływ „obszaru" nie jest modyfikowany przez „wiek".

Rozważany przykład przedstawia model statystyczny przydatny do przeprowadzenia analizy regresji Poissona przy dwóch czynnikach. W ogólności, zamiast dwóch czynników (wiek i obszar), możemy mieć p - czynników: X_1, X_2, ..., X_p. Wtedy ogólna metoda dopasowywania modelu regresji Poissona nie zmienia się i polega na wykorzystaniu rozkładu Poissona do otrzymania funkcji wiarygodności, która może być później maksymalizowana w celu otrzymania estymatorów parametrów modelu oraz oszacowanych błędów standardowych zmaksymalizowanych statystyk MNW. Ponieważ pakiety programów (zawarte np. w systemie analiz statystycznych SAS) mogą wykonywać takie analizy, zatem użytkownik musi jedynie wyszczególnić trafny model, który ma być dopasowany. Numeryczna analiza dla powyższego przykładu zostanie przeprowadzona w dalszej części.

D1.9 Ogólna analiza regresja Poissona w MNW

Kontynuujmy opis ogólnej struktury analizy regresji Poissona, tym razem z punktu widzenia estymacji parametrów MNW. Zmienną objaśnianą Y jest liczba zliczeń defektów (silników) otrzymanych dla każdej podgrupy, której wartości są wyjaśniane w regresji przez ustaloną liczbę czynników X_1, X_2, ... , X_p .

Rozważmy wstępnie indeksowanie grupy jednym indeksem. Dla grupy j , gdzie j = 1,2, ... ,N, przez Y_j oznaczmy zmienną obserwowanej ilości defektów oraz przez ℓ_j, całkowitą wielkość zakumulowanego czasu dla wszystkich samochodów w j-tej podgrupie.

D1.9.1 Tempo defektów

Oznaczmy przez $\vec{X}_j = \left(X_{1j}, X_{2j}, ..., X_{pj}\right)$ zbiór zmiennych X_1, X_2, ... , X_p, dla j-tej podgrupy. Przez $\beta = \left(\beta_0, \beta_1, \beta_2, ..., \beta_p\right)$ oznaczmy zbiór nieznanych parametrów, a przez $r\left(\vec{X}_j, \beta\right)$ funkcję czynników $\vec{X}_j$ oraz tych parametrów, określoną w regresji Poissona [1] następująco:

$$r\left(\vec{X}_j, \beta\right) = \exp\left(\beta_0 + \sum_{i=1}^{p} \beta_i X_{ij}\right), \qquad j = 1, 2, \ldots, N, \tag{16}$$

opisującą ryzyko awarii, czyli liczbę defektów w jednostce czasu (tempo defektów) dla *j*-tej podgrupy. Inne postacie funkcji $r\left(\vec{X}_j, \beta\right)$ można znaleźć w [4].

Zatem zgodnie z poprzednimi rozważaniami, funkcja $r\left(\vec{X}_j, \beta\right)$ jest miarą wskazującą ile defektów zdarza się w jednostce czasu, czyli $r\left(\vec{X}_j, \beta\right)$ mierzy tempo pojawiania się defektów w jednostce czasu badania.

D1.9.2 Oczekiwana liczbę zdarzeń

Zgodnie z postacią funkcji regresji (6), możemy w *j*-tej podgrupie, gdzie *j*=1,2,...,*N*, i pod warunkiem $r\left(\vec{X}_j, \beta\right) > 0$, zapisać **oczekiwaną warunkową liczbę zdarzeń** następująco [1]:

$$E\left(Y_j\right) = \mu_j = \ell_j \, r\left(\vec{X}_j, \beta\right) \quad , \qquad j = 1, 2, \ldots, N \, , \tag{17}$$

gdzie Y_i jest zmienną losową Poissona. Równanie (17) jest treścią równania regresji pierwszego rodzaju. Jego współczynniki musimy oszacować na podstawie pobranej próbki. Równanie regresji z oszacowanymi współczynnikami nazywamy *równaniem regresji drugiego rodzaju*.

Ponieważ z założenia Y_i ma rozkład Poissona ze średnią μ_i, a więc:

$$p\left(y_j \mid \mu_j\right) = \frac{\mu_j{}^{y_j} e^{-\mu_j}}{y_j!} \qquad , \qquad j = 1, 2, \ldots, N \, . \tag{18}$$

Tak więc z równań (17) oraz (18) wynika [4]:

$$p\left(y_j \mid \beta\right) = \frac{\left[\ell_j \, r\left(\vec{X}_j, \beta\right)\right]^{y_j} e^{-\ell_j r\left(\vec{X}_j, \beta\right)}}{y_j!} \, , \quad \text{gdzie} \quad y_j = 0, 1, \ldots, \infty \quad \text{oraz} \quad j = 1,2, \ldots, N \, . \tag{19}$$

Podsumowanie: Widzimy, że analiza regresji dotyczy modelowania (warunkowej) wartości oczekiwanej zmiennej zależnej, tzn. objaśnianej, jako funkcji określonych czynników $\vec{X}_j$. Forma funkcji wiarygodności, która jest użyta do estymacji zbioru współczynników regresji β, jest zależna od założeń uczynionych o postaci rozkładu zmiennej objaśnianej *Y*.

D1.10 Funkcja wiarygodności dla analizy regresji Poissona

Zgodnie z założeniami MNW [1], w celu otrzymania funkcji wiarygodności przyjmijmy, że próba $\tilde{Y} = (Y_1, Y_2, ..., Y_N)$ jest układem *N* wzajemnie niezależnych zmiennych losowych. W przypadku gdy *Y* ma rozkład Poissona zmienne Y_j, *j*=1,2,...,*N*, też mają rozkład Poissona (19). Zatem funkcja wiarygodności dla analizy regresji Poissona ma ogólną postać [4, 1]:

$$P\left(\tilde{Y} \mid \beta\right) = \prod_{j=1}^{N} p\left(Y_j \mid \beta\right) = \prod_{j=1}^{N} \left\{ \frac{\left[\ell_j r\left(\vec{X}_j, \beta\right)\right]^{Y_j} e^{\ell_j r\left(\vec{X}_j, \beta\right)}}{Y_j!} \right\} =$$

$$= \frac{\left\{ \prod_{j=1}^{N} \left[\ell_j r\left(\vec{X}_j, \beta\right)\right]^{Y_j} \right\} \exp\left[-\sum_{j=1}^{N} \ell_j r\left(\vec{X}_j, \beta\right) \right]}{\prod_{j=1}^{N} Y_j!} \qquad (20)$$

gdzie $E\left(Y_j\right) = \mu_j = \ell_j r\left(\vec{X}_j, \beta\right)$, j = 1, 2, ... , *N*. Liczba parametrów (β_i) w funkcji wiarygodności (20) wynosi $p+1$.

Aby móc w praktyce wykorzystać postać (20) funkcji wiarygodności, musimy podać konkretną postać funkcji tempa $r\left(\vec{X}_j, \beta\right)$. Specyfikacja ta musi być dostosowana do typu procesu i uprzedniej wiedzy oraz doświadczenia z badaniem związków pomiędzy rozważanymi zmiennymi. Przykłady możliwych wyborów dla $r\left(\vec{X}_j, \beta\right)$ można znaleźć w [6, 4, 1]. W regresji Poissona ma ona postać:

$$r\left(\vec{X}_j, \beta\right) = \exp(\lambda_j^*) \ , \quad \text{gdzie} \quad \lambda_j^{\ *} = \beta_0 + \sum_{i=1}^{p} \beta_i X_{ij} \ , \quad \text{dla} \ \ j = 1,2,...,N. \qquad (21)$$

D1.10.1 Uwaga o regresji Poissona i liniowej regresji wielorakiej

Jedyną modelową różnicą pomiędzy regresją Poissona a standardową regresją wieloraką jest to, że pierwsza zakłada zastosowanie rozkładu Poissona, podczas gdy druga zakłada zastosowanie rozkładu normalnego. W każdym przypadku cel analizy jest jednak taki sam, tzn. dopasowanie (dofitowanie) do danych równania regresji, które będzie trafnie modelowało

E(Y_j) jako funkcję czynników X_1, X_2, ... , X_p, co jest sednem równania regresji Poissona zadanego poniższymi wzorami (16)-(17).

W modelu regresji wielorakiej stosujemy funkcję regresji typu $E(Y_j) = \beta_0 + \sum_{i=1}^{p} \beta_i X_{ij}$. Jeśli zamiast niej zastosujemy $E(Y_j) = \exp(\beta_0 + \sum_{i=1}^{p} \beta_i X_{ij})$, używaną w regresji Poissona, to postać modelu staje nieliniowa i przechodzimy od liniowej do nieliniowej analizy regresji. Główną konsekwencją powyższej zmiany jest to, że zamiast równań wiarygodności liniowych musimy rozwiązać zbiór równań nieliniowych dla współczynników β. Rozwiązanie to wymaga zazwyczaj zastosowania pewnego rodzaju procedury iteracyjnej wspomaganej komputerowo.

D1.11 Równania wiarygodności oraz IRLS

Estymatory MNW $\hat{\beta}_0$, $\hat{\beta}_1$, ... , $\hat{\beta}_p$ parametrów β_0, β_1, ... , β_p są otrzymywane dla funkcji wiarygodności określonej w (20) jako efekt rozwiązania p +1 równań wiarygodności:

$$\frac{\partial}{\partial \beta_i}\left[\ln P(\tilde{Y}, \beta)\right] = 0 \qquad , \qquad i = 0, 1, \ldots , p. \tag{22}$$

Rozwiązanie układu równań wiarygodności (22) jest zazwyczaj otrzymane pewnym iteracyjnym algorytmem komputerowym wykorzystującym np. metodę numeryczną Newtona-Raphson'a [6]. Fakt, że zachodzi $E(Y_j) = \sigma^2(Y_j) = \mu_j = \ell_j r(\vec{X}_j, \beta)$ oznacza, że wariancja zmiennej objaśnianej nie jest stała, ale zmienia się jako funkcja parametru ℓ_j oraz $\vec{X}_j$. Ponieważ wariancja $\sigma^2(Y_j)$ jest również funkcją zbioru parametrów β, zatem wagi w tego typu analizie regresji zmieniają się z powodu zmiany wartości estymatora $\hat{\beta}$ w każdym kroku procesu iteracji i w każdym iteracyjnym kroku wymagane jest ich ponowne przeliczenie. Najpopularniejszym z algorytmów jest algorytm nazywany metodą iteracyjnie przeliczonych wag najmniejszych kwadratów (*iteratively reweighted least squares* (IRLS)) [6, 4]. Nazwa ta nie odnosi się do MNK, która ma probabilistyczne znaczenie tylko gdy zmienne Y_j mają rozkład normalny.

Kilka statystycznych pakietów, takich jak SAS, używający procedury GENMOD lub NLMIXED, może być wykorzystywanych w celu znalezienia estymatorów MNW $\hat{\beta}$ parametrów β występujących w funkcji wiarygodności (20).

D1.12 Macierz kowariancji i obserwowana informacja Fishera

Dodatkowo procedurami SAS, estymowana jest obserwowana macierz kowariancji $\hat{V}(\hat{\beta})$ estymatorów parametrów β będąca w metodzie MNW odwrotnością *obserwowanej* informacji Fishera iF :

$$\hat{\mathrm{V}}(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ...,) = \begin{bmatrix} \hat{\sigma}^2(\hat{\beta}_0) & \hat{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \hat{Cov}(\hat{\beta}_0, \hat{\beta}_2) \\ \hat{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \hat{\sigma}^2(\hat{\beta}_1) & \hat{Cov}(\hat{\beta}_1, \hat{\beta}_2) \\ \hat{Cov}(\hat{\beta}_0, \hat{\beta}_2) & \hat{Cov}(\hat{\beta}_1, \hat{\beta}_2) & \hat{\sigma}^2(\hat{\beta}_2) \\ & & \ddots \end{bmatrix} := iF^{-1} \quad , \qquad (23)$$

oraz miary dobroci dopasowania rozważanego modelu i pewne statystyki diagnostyczne regresji, użyteczne dla wykrywania obserwacji wpływowych oraz współliniowości [4]. Wszystko to w raportach SASa pojawia się jako część wydruku komputerowego. Więcej na temat *obserwowanej* informacji Fishera iF , jej definicji oraz przykładów, można znaleźć w [5, 1].

D1.13 Analiza regresji dla przykładu: Model 1

Pierwszy rozważany model regresji Poissona dla oczekiwanej liczby przypadków awarii silnika w podgrupach (i, j) ma postać zadaną przez (6) oraz-(9). Jest więc to uprzednio wprowadzony Model 1:

$$E(Y_{ij}) = \mu_{ij} = \ell_{ij} r_{ij} \quad , \qquad i = 1, 2, \ldots, 8 ; \quad j = 0, 1 , \qquad (24)$$

gdzie:

Model 1: $$\ln r_{ij} = \alpha + \sum_{k=1}^{7} \alpha_k U_k + \beta M \quad . \qquad (25)$$

Zmienne U_k były „sztucznie” wprowadzonymi zmiennymi kierunkowymi (ukrytymi) (7) wskazującymi wariant wiekowy i przyjmującymi wartości 0 lub 1, a zmienna kierunkowa M przyjmowała zgodnie z (8) wartości 0 lub 1, wskazując odpowiednio obszar Miejski lub Wiejski.

Dla powyższego modelu ryzyko względne wynosi (4):

$$R_{Wi} = \frac{r_{i1}}{r_{i0}} \quad , \tag{26}$$

a zgodnie z (14) jego postać redukuje się do:

$$R_{Wi} = e^{\beta} \quad , \tag{27}$$

gdzie e^{β} jest niezależne od i, reprezentując ogólne ryzyko względne dopasowane do „wieku".

Konkretna postać funkcji wiarygodności powyższego modelu jest konsekwencją założenia, że liczba zliczeń Y_{ij} ma rozkład Poissona ze średnią $\mu_{ij} = \ell_{ij} r_{ij}$. Zgodnie z (20) ma ona w próbce postać:

$$P\left(\vec{y} \mid \beta\right) = \prod_{i=1}^{8} \left\{ \left[\frac{\left(\ell_{i0} r_{i0}\right)^{y_{i0}} e^{-\ell_{i0} r_{i0}}}{y_{i0}!} \right] \left[\frac{\left(\ell_{i1} r_{i1}\right)^{y_{i1}} e^{-\ell_{i1} r_{i1}}}{y_{i1}!} \right] \right\} \quad , \tag{28}$$

gdzie zgodnie z (10)-(11) mamy $r_{i0} = \exp\left(\alpha + \alpha_i\right)$ i $r_{i1} = \exp\left(\alpha + \alpha_i + \beta\right)$ dla i = 1, ..., 7 , oraz $r_{80} = \exp\left(\alpha\right)$ i $r_{81} = \exp\left(\alpha + \beta\right)$ dla i = 8 .

Użycie pakietu komputerowego dla regresji Poissona będzie maksymalizowało powyższą funkcję wiarygodności, dając 9 estymatorów parametrów badanego modelu:

$$\left\{\hat{\alpha}, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\alpha}_4, \hat{\alpha}_5, \hat{\alpha}_6, \hat{\alpha}_7, \hat{\beta}\right\} \tag{29}$$

oraz oszacowaną 9×9 - wymiarową macierz kowariancji (23).

D1.14 Analiza numeryczna programem SAS

W celu wykonania selekcji modelu regresji Poissona dla powyższego przykładu z wykorzystaniem SAS należy utworzyć zbiór danych oraz program wyznaczający oszacowania parametrów modelu zgodnie z procedurą języka programowania 4GL tej aplikacji. Następnie zbiór danych należy umieścić w edytorze systemu SAS (Widok -> Enhanced Editor) i uruchamiając właściwą procedurę, dokonać przeliczenia modelu (Uruchom -> Przekaż) [8]. Język 4GL dzięki swojej budowie umożliwia przetwarzanie oraz pełną obsługę zbiorów danych. Ramka ogólnej składni wprowadzonej procedury ma postać:

PROC nazwa procedury DATA=zbior_danych opcje_procedury;

...

Instrukcje;

...

RUN;

Niektóre z wykorzystywanych poniżej instrukcji, potrzebnych w dalszej analizie, podane zostały w Uzupełnieniu. Pełny ich wykaz oraz zastosowanie można znaleźć w pomocy pakietu SAS.

D1.14.1 Dane oraz programy

W analizie rozważanego przykładu można wykorzystać jeden, poniższy zbiór danych, wprowadzając w zależności od modelu odpowiednie modyfikacje dopiero na poziomie programu analizującego rozważany model. Wyjaśnienie używanych poleceń języka 4GL oraz opis zmiennych A do O znajdują się w Uzupełnieniu.

Zbiór danych:

```
data model;
input A Y  N  M  U1  U2  U3  U4  U5  U6  U7 U1M U2M U3M U4M U5M U6M U7M O;
ln = log(N);
datalines;
1    1    172675    0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
2    16   123065    0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0
3    30   96216     0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0
4    71   92051     0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0
5    102  72159     0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0
6    130  54722     0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0
7    133  32185     0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0
8    40   8328      0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
1    4    181343    1  1  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0
2    38   146207    1  0  1  0  0  0  0  0  0  1  0  0  0  0  0  0  0
3    119  121374    1  0  0  1  0  0  0  0  0  0  1  0  0  0  0  0  0
4    221  111353    1  0  0  0  1  0  0  0  0  0  0  1  0  0  0  0  0
5    259  83004     1  0  0  0  0  1  0  0  0  0  0  0  1  0  0  0  0
6    310  55932     1  0  0  0  0  0  1  0  0  0  0  0  0  1  0  0  0
7    226  29007     1  0  0  0  0  0  0  1  0  0  0  0  0  0  1  0  0
8    65   7538      1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
;
run;
```

Wczytanie programu analizującego model:

Po wczytaniu zbioru danych, przystępujemy do wpisania programu analizującego konkretną postać modelu, który wykorzysta całość lub część powyższego zbioru danych, w zależności od modelu regresji Poissona dla przykładu awarii silnika. I tak dla Modelu 1 program ten, wykorzystujący procedurę GENMOD, ma następującą postać:

proc ***genmod*** *data = model;*
model Y = M U1 U2 U3 U4 U5 U6 U7 / covb
dist = poisson
link = log
offset = ln*;*
run;
quit;

Uwaga:
Zamiast wpisywać *model* w powyższej postaci można użyć wprowadzonej zmiennej klasującej A i zamienić odpowiednie linie:

ref = 8;
class A;
model Y = A / cov

a program SAS odda raport identycznej postaci.

D1.14.2 Wynik analizy numerycznej SAS dla Modelu 1

Jako rezultat wczytania powyższych danych i uruchomienia procedury GENMOD dla rozważanego aktualnie Modelu 1, otrzymujemy raport systemu SAS, który ma następującą postać:

System SAS
The GENMOD Procedure

Informacje o modelu

Zbiór	WORK.MODEL1
Rozkład	Poisson
Funkcja wiążąca	Log
Zmienna zależna	Y
Zmienna przesunięcia	ln
Liczba obserwacji wczytanych	16
Liczba obserwacji użytych	16

Informacje o poziomie klasyfikacji

Klasa	Poziomy	Wartości
A	8	1 2 3 4 5 6 7 8

Informacje o parametrach

Parametr	Efekt
Prm1	Intercept
Prm2	M
Prm3	U1
Prm4	U2
Prm5	U3
Prm6	U4
Prm7	U5
Prm8	U6
Prm9	U7

Kryteria oceny zgodności

Kryterium	St. sw.	Wartość	Wartość/st. sw.
Dewiancja	7	**8.1950**	1.1707
Skalowana dewia	7	8.1950	1.1707
Chi-kwadrat Pearso	7	8.0626	1.1518
Scaled Pearson X2	7	8.0626	1.1518
Log. wiarogodn		7201.8635	

System SAS
The GENMOD Procedure

Algorytm osiągnął zbieżność.

Macierz kowariancji szacunkowych

	Prm1	Prm2	Prm3	Prm4	Prm5
Prm1	0.01074	-0.001824	-0.009465	-0.009419	-0.009398
Prm2	-0.001824	0.002725	-0.000087	-0.000156	-0.000188
Prm3	-0.009465	-0.000087	0.20953	0.009529	0.009530
Prm4	-0.009419	-0.000156	0.009529	0.02805	0.009535
Prm5	-0.009398	-0.000188	0.009530	0.009535	0.01625
Prm6	-0.009413	-0.000166	0.009529	0.009533	0.009535
Prm7	-0.009431	-0.000138	0.009528	0.009532	0.009533
Prm8	-0.009476	-0.000072	0.009526	0.009528	0.009529
Prm9	-0.009526	2.5943E-6	0.009524	0.009524	0.009524

Macierz kowariancji szacunkowych

	Prm6	Prm7	Prm8	Prm9
Prm1	-0.009413	-0.009431	-0.009476	-0.009526
Prm2	-0.000166	-0.000138	-0.000072	2.5943E-6
Prm3	0.009529	0.009528	0.009526	0.009524
Prm4	0.009533	0.009532	0.009528	0.009524
Prm5	0.009535	0.009533	0.009529	0.009524
Prm6	0.01296	0.009532	0.009528	0.009524
Prm7	0.009532	0.01230	0.009527	0.009524
Prm8	0.009528	0.009527	0.01180	0.009524
Prm9	0.009524	0.009524	0.009524	0.01231

Analiza ocen parametrów

Parametr	St. sw.	Ocena	Błąd standardowy	95% granice przedziału ufności Walda		Chi-kwadrat	Pr > chi kw..
Intercept	1	-5.4797	0.1037	-5.6828	-5.2765	2794.67	<.0001
M	1	**0.8043**	**0.0522**	0.7020	0.9066	237.34	<.0001
U1	1	-6.1782	0.4577	-7.0753	-5.2810	182.17	<.0001
U2	1	-3.5480	0.1675	-3.8763	-3.2197	448.76	<.0001
U3	1	-2.3308	0.1275	-2.5807	-2.0810	334.36	<.0001
U4	1	-1.5830	0.1138	-1.8061	-1.3599	193.38	<.0001
U5	1	-1.0909	0.1109	-1.3083	-0.8735	96.75	<.0001
U6	1	-0.5328	0.1086	-0.7457	-0.3199	24.06	<.0001
U7	1	-0.1196	0.1109	-0.3371	0.0978	1.16	0.2809
Skala	0	1.0000	0.0000	1.0000	1.0000		

UWAGA: The scale parameter was held fixed.

D1.14.3 Oszacowanie parametru i błąd standardowy oszacowania dla Modelu 1

Z powyższego raportu możemy odczytać *oszacowanie* $\hat{\beta}$ *MNW parametru* β:

$$\hat{\beta} = 0{,}8043 \tag{30}$$

oraz *błąd standardowy* (*s.e.*) tego oszacowania wyznaczony jako fragment macierzy (wariancji-) kowariancji (23) [4]:

$$\hat{\sigma}_{\hat{\beta}} = 0{,}0522 \ . \tag{31}$$

Punktowe oszacowanie $\hat{r}_{Wi}^{Model1}$ dopasowanego ze względu na wiek ryzyka względnego R_{Wi} wynosi więc:

$$\hat{r}_{Wi}^{Model1} = e^{\hat{\beta}} = e^{0{,}8043} = 2{,}23513 \ . \tag{32}$$

Natomiast 95%-owy wiarygodnościowy przedział ufności dla e^{β} [4], przy odwołaniu się do faktu, że dla dużej próbki estymator MNW ma *w przybliżeniu* rozkład normalny, ma postać:

$$\exp[\hat{\beta} \pm 1{,}96\ \hat{\sigma}_{\hat{\beta}}] = \exp\left[0{,}8043 \pm 1{,}96\ (0{,}0522)\right] = \exp(0{,}8043 \pm 0{,}1023) \ , \tag{33}$$

lub

$$\left(e^{0.7020}, e^{0.9066}\right) = (2.01778;\ 2{,}47589) \ . \tag{34}$$

D1.14.4 Test hipotezy zerowej i statystyka Wald'a

Dla dużej próbki, test hipotezy zerowej:

$$H_0: \beta = 0 \tag{35}$$

o braku zależności korelacyjnej tempa zachorowań od lokalizacji, wobec hipotezy alternatywnej:

$$H_1: \beta \neq 0 \ , \tag{36}$$

może być przeprowadzony z zastosowaniem statystyki Wald'a [4]:

$$U = \frac{\hat{\beta} - 0}{\hat{\sigma}_{\hat{\beta}}} \quad . \tag{37}$$

Przy prawdziwości hipotezy zerowej H_0: $\beta = 0$ ma ona asymptotycznie rozkład normalny N(0,1).

Dla rozważanego przykładu wartość statystyki Wald'a wynosi:

$$U = \frac{0{,}8043 - 0}{0{,}0522} = 15{,}408 \ , \tag{38}$$

natomiast empiryczny poziom istotności [4, 1] ma wartość (wyznaczoną np. w pakiecie kalkulacyjnym Excel):

$$p = \Pr\left(|U| \geq 15{,}408\right) < 0{,}0001 \ . \tag{39}$$

D1.14.5 Wniosek

Ze względu na $p < 0{,}0001$ przeprowadzona analiza regresji Poissona wskazuje na statystycznie istotny wpływ lokalizacji (tzn. na statystyczną istotność wprowadzenia parametru β przy zmiennej kierunkowej *M* wskazującej lokalizację). Ze względu na wartość oszacowanego ryzyka względnego $\hat{r}_{Wi} = e^{\hat{\beta}} = e^{0{,}8043} = 2{,}23513$ ogólne, dopasowane ze względu na wiek, tempo awarii silników samochodów na Wsiach jest około 2,2 razy większe niż w Miastach. Wyznaczony 95%-owy przedział ufności dla ogólnego dopasowania ryzyka względnego wynosi $(2{,}01776; 2{,}47591)$.

Do analizy Modelu 1 wrócimy jeszcze poniżej, aby omówić interakcję czynnika „wiek" ze zmienną „obszar", bądź uwzględnienie „wieku" jako ewentualnego zaburzenia w modelu [4] oraz porównać dobroć dopasowania Modelu 1 z innymi modelami w hierarchii.

D1.15 Miary dobroci dopasowania

Do weryfikacji hipotez o nie występowaniu statystycznie istotnego braku dopasowania w jednym modelu w porównaniu z innym modelem, będącym członkiem tej samej hierarchii modeli, wykorzystamy logarytmiczny ilorazu zmaksymalizowanych wiarygodności tych modeli oraz dewiancję, jako jego szczególny typ. Tak jak dotychczas, zmienne Y_1, Y_2, ..., Y_N w próbie są wzajemnie niezależne, a każda z nich ma rozkład Poissona (18).

D1.15.1 Wiarygodność modelu podstawowego

Nie wprowadzając zależności od czynników X_1, X_2, ... , X_p, jedynymi parametrami, które wchodzą w funkcję wiarygodności są wartości oczekiwane $\mu_1, \mu_2, \ldots, \mu_N$ zmiennych $Y_1, Y_2, \ldots, Y_N$. Dlatego wiarygodność próby $\tilde{Y}$ przyjmuje dla „**modelu podstawowego**" postać [4, 1]:

$$P\left(\tilde{Y} \mid \mu\right) = \prod_{j=1}^{N} \frac{{\mu_j}^{Y_j} e^{-\mu_j}}{Y_j!} = \frac{\left(\prod_{j=1}^{N} {\mu_j}^{Y_j}\right) \exp\left(-\sum_{j=1}^{N} \mu_j\right)}{\prod_{j=1}^{N} Y_j!} \quad , \tag{40}$$

gdzie $\mu = (\mu_1, \mu_2, \ldots, \mu_N)$. Zatem N jest równocześnie liczebnością zbioru danych, która może być liczbą podgrup, komórek lub kategorii, oraz liczbą parametrów modelu podstawowego występującą w wiarygodności (40).

D1.16 Układ równań wiarygodności modelu podstawowego

Ponieważ układ równań wiarygodności:

$$\frac{\partial}{\partial \mu_j}\left[\ln P\left(\tilde{Y} \mid \mu\right)\right] = 0 \qquad , \qquad j = 1, 2, ..., N \tag{41}$$

ma dla modelu podstawowego (40) rozwiązanie:

$$\hat{\mu} = (\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_N) = (Y_1, Y_2, \ldots, Y_N) \ , \tag{42}$$

zatem zmaksymalizowana funkcja wiarygodności (40) przyjmuje dla modelu podstawowego postać:

$$P\left(\tilde{Y} \mid \hat{\mu}\right) = \frac{\left(\prod_{j=1}^{N} Y_j^{\,Y_j}\right)\exp\left(-\sum_{j=1}^{N} Y_j\right)}{\prod_{j=1}^{N} Y_j!} \quad . \tag{43}$$

O ile liczba parametrów modelu spełnia warunek $(p+1) < N$, to wartość $P(\vec{y} \mid \hat{\mu})$ zmaksymalizowanej wiarygodności $P\left(\tilde{Y} \mid \hat{\mu}\right)$, (43), wyznaczona dla dowolnych danych jest większa niż ta, którą otrzymalibyśmy przez zmaksymalizowanie funkcji wiarygodności (20) związane z rozwiązania równań wiarygodności (22). Jest tak, ponieważ (40) nie nakłada ograniczeń na strukturę parametru μ_i, podczas gdy (20) nakłada ograniczenie postaci $\mu_j = \ell_j r\left(\vec{X}_j, \beta\right)$, które siłą rzeczy dla $(p+1) < N$ pogarsza dopasowanie modelu do danych pomiarowych w porównaniu z modelem podstawowym.

Pomyśl o tym tak: Model podstawowy dopasowuje się do danych, w każdym punkcie z osobna, leżąc zgodnie z (42) maksymalnie blisko tych danych, natomiast MNW dla modelu regresji $E\left(Y_j\right) = \mu_j = \ell_j\, r\left(\vec{X}_j, \beta\right)$, (17), wyznacza krzywą regresji przechodzącą pomiędzy punkami pomiarowymi.

Powód konstrukcji modelu regresji: Powodem analizowania modelu regresji, a nie trwania przy modelu podstawowym, nie jest sama dokładność dopasowania, lecz próba zrozumienia istoty opisywanego zjawiska oraz mniejsza liczba parametrów, co wpływa na zmniejszenie kosztów oszacowywania parametrów z określoną dokładności [4].

D1.17 Postać funkcji wiarygodności dla hipotezy zerowej modelu

Funkcja wiarygodności (20) może być rozumiana, jako ta, którą należy się posłużyć, gdy prawdziwa jest następująca hipoteza zerowa:

$$\mathrm{H}_0\colon\ \mu_j = \ell_j r\left(\vec{X}_j, \beta\right), \qquad j = 1, 2, ..., N, \tag{44}$$

podczas gdy (40) jest funkcją wiarygodności przy prawdziwości hipotezy alternatywnej:

$$\mathrm{H}_1\text{: nie ma ograniczenia struktury } \mu_j\ , \quad j = 1, 2, ..., N\ . \tag{45}$$

D1.18 Dewiancja

Jeżeli $P\left(\tilde{Y} \mid \hat{\beta}\right)$ jest zmaksymalizowanym prawdopodobieństwem przy wiarygodności (20), gdzie $\hat{\beta}$ jest zbiorem estymatorów MNW parametrów β, wtedy statystyka ilorazu wiarygodności nazywana dewiancją i określona następująco [4, 1]:

$$D\left(\hat{\beta}\right) = -2\ln\left[\frac{P\left(\tilde{Y} \mid \hat{\beta}\right)}{P\left(\tilde{Y} \mid \hat{\mu}\right)}\right] \quad , \tag{46}$$

służy do oceny, czy przy ustalonych danych pomiarowych, wartość wiarygodności $P\left(\vec{y} \mid \hat{\beta}\right)$ dla modelu $\mu_j = \ell_j r\left(\vec{X}_j \mid \beta\right)$ jest istotnie statystycznie mniejsza niż $P\left(\vec{y} \mid \hat{\mu}\right)$ **modelu podstawowego, który nie narzuca struktury na** μ_j.

W przypadku wpadnięcia wartości statystyki (46) w (wiarygodnościowy) obszar krytyczny [1], test statystyczny wskazuje na statystycznie istotny brak dopasowania do danych postawionego w hipotezie zerowej (44) modelu regresji $\mu_j = \ell_j r\left(\vec{X}_j \mid \beta\right)$ [4, 1].

Uwaga: Dewiancja (reszt) może być rozumiana jako miara zmienności wartości y_j wokół dopasowanego modelu, na którym leżą wartości przewidywane $\hat{y}_j$ przez model [1].

D1.18.1 Minimalny oszczędny model opisu danych

Model podstawowy bez struktury parametrów zawiera tyle parametrów ile jest grup danych pomiarowych, czyli N. Celem analizy regresji jest otrzymanie oszczędnego opisu danych. Model $\mu_j = \ell_j r\left(\vec{X}_j \mid \beta\right)$, zawierający $p+1$ parametrów, uznamy za oszczędny, jeśli ma wartość zmaksymalizowaną wiarygodności prawie tak dużą, jak pojawiająca się dla modelu podstawowego i jednocześnie najmniejszą liczbę parametrów funkcji regresji w klasie modeli hierarchicznych, do których należy. Dla modelu oszczędnego wartość dewiancji wpadnie w wiarygodnościowy obszar przyjęć hipotezy zerowej.

D1.18.2 Rozkład statystyki dewiancji

Dla *bardzo dużej* próbki dewiancja $D\left(\hat{\beta}\right)$ posiada, przy prawdziwości hipotezy $H_0: \mu_j = \ell_j r\left(\vec{X}_j \mid \beta\right)$, (44), w przybliżeniu rozkład chi-kwadrat z $N - p - 1$ stopniami swobody [1]. Zatem, przybliżony test dobroci dopasowania do danych modelu $\mu_j = \ell_j r\left(\vec{X}_j \mid \beta\right)$, może zostać wykonany przez sprawdzenie czy wyznaczona w próbce wartości $D\left(\hat{\beta}\right)$ jest nie mniejsza niż wartość krytyczna w prawym ogonie rozkładu chi-kwadrat z $N - p - 1$ stopniami swobody [4].

Alternatywnie, mając wartości $D\left(\hat{\beta}\right)$, można policzyć empiryczny poziom istotności $p = \Pr(\chi^2_{N-p-1} \geq D(\hat{\beta}))$ i porównać jego wartość z przyjętą (w dziedzinie badań) wartością poziomu istotności α [9]. Gdy $p \leq \alpha$ wtedy odrzucamy hipotezę zerową H_0, mówiącą o nie występowaniu braku dopasowania w badanym modelu regresji w porównaniu z modelem podstawowym i decydujemy się na statystycznie uzasadnioną rozbudowę modelu, o dalsze parametry strukturalne.

D1.18.3 Dewiancja, odpowiedź układu przewidywana modelem i wartości pomiarowe

Jeśli przez $\mu_j = \ell_j r\left(\vec{X}_j \mid \beta\right)$ oznaczymy przewidywaną odpowiedź układu, tzn. przewidywanie modelu regresji (17) dla Y_j w *j*-tej komórce, wtedy wielkość (46) może być zapisana w postaci:

$$D\left(\hat{\beta}\right) = 2\sum_{j=1}^{N}\left[Y_j \ln\left(\frac{Y_j}{\hat{Y}_j}\right) - \left(Y_j - \hat{Y}_j\right)\right] \quad . \tag{47}$$

Dewiancja $D\left(\hat{\beta}\right)$ zachowuje się podobnie jak suma kwadratów reszt $SKR = \sum_{j=1}^{N}\left(Y_j - \hat{Y}_j\right)^2$ w standardowej analizie liniowej regresji wielorakiej [4,1]. Kiedy dopasowywany model dokładnie prognozuje obserwowane dane, tzn. $Y_j = \hat{Y}_j$, j = 1, 2, ..., N, wtedy $D\left(\hat{\beta}\right) = 0$. Natomiast im większa rozbieżność pomiędzy obserwowanymi i prognozowanymi danymi, tym większa jest wartość $D\left(\hat{\beta}\right)$.

Kiedy wszystkie prognozowane wartości mają rozsądną wielkość (tzn. $\hat{Y}_j > 3$, j = 1, 2, ..., N), wtedy (47) można dla dużej próby przybliżyć znaną statystyką chi-kwadrat Pearson'a, która ma postać [4]:

$$\chi^2 = \sum_{j=1}^{N} \frac{\left(Y_j - \hat{Y}_j\right)^2}{\hat{Y}j} \quad , \tag{48}$$

gdzie Y_j jest wartością obserwowaną, a $\hat{Y}_j$ wartością przewidywaną. Jednak statystyka (48) może przyjmować myląco duże wartości, gdy pewne $\hat{Y}_j$ są bardzo małe.

D1.19 Porównanie dwóch modeli z wykorzystaniem dewiancji

Dewiancje dla różnych modeli z jednej hierarchicznej klasy modeli mogą być użyte do tworzenia statystyk wiarygodności. Rozważamy funkcję wiarygodności (20), zawierającą zbiór parametrów $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$, z dewiancją $D(\hat{\beta})$ daną równaniem (46).

Załóżmy, że $0 < r < p$ i przeprowadźmy statystyczny test hipotezy mówiącej, że ostatnich $p - r$ parametrów w β jest równych zero.

Rozważana hipoteza zerowa ma więc postać:

$$\mathrm{H}_0\colon\ \beta_{r+1} = \beta_{r+2} = \ldots = \beta_p = 0 \ , \tag{49}$$

a hipoteza alternatywna:

H_1 postuluje, że *przynajmniej jeden z parametrów* $\beta_{r+1}, \beta_{r+2}, \ldots, \beta_p$ *jest różny od zera*.

Przy prawdziwości hipotezy H_0, funkcję wiarygodności modelu niższego można otrzymać zastępując β w (20) przez $\beta_{(r)}$, gdzie:

$$\beta_{(r)} \equiv (\beta_0, \beta_1, \ldots, \beta_r; 0,0,\ldots,0) \ . \tag{50}$$

Jeśli funkcję wiarygodności z parametrami $\beta_{(r)}$ oznaczymy jako $P\left(\tilde{Y} \mid \beta_{(r)}\right)$ i jeśli $\hat{\beta}_{(r)}$ jest estymatorem MNW parametru $\beta_{(r)}$, to po wstawieniu $\hat{\beta}_{(r)}$ do tej funkcji wiarygodności, test ilorazu wiarygodności dla hipotezy zerowej H_0 jest wykonywany z użyciem następującej statystyki testowej logarytmu ilorazu wiarygodności:

$$-2\ln\left[\frac{P\left(\tilde{Y} \mid \hat{\beta}_{(r)}\right)}{P\left(\tilde{Y} \mid \hat{\beta}\right)}\right], \tag{51}$$

która przy założeniu, że hipoteza H_0: $\beta_{r+1} = \beta_{r+2} = \ldots = \beta_p = 0$ jest prawdziwa i dla dużej próbki ma w przybliżeniu rozkład chi-kwadrat z $(N-r-1)-(N-p-1) = p-r$ stopniami swobody.

Łatwo sprawdzić, że wielkość (51) jest równa różnicy dewiancji:

$$D\left(\hat{\beta}_{(r)}\right) - D\left(\hat{\beta}\right) = -2\ln\left[\frac{P\left(\tilde{Y} \mid \hat{\beta}_{(r)}\right)}{P\left(\tilde{Y} \mid \hat{\mu}\right)}\right] + 2\ln\left[\frac{P\left(\tilde{Y} \mid \hat{\beta}\right)}{P\left(\tilde{Y} \mid \hat{\mu}\right)}\right] = -2\ln\left[\frac{P\left(\tilde{Y} \mid \hat{\beta}_{(r)}\right)}{P\left(\tilde{Y} \mid \hat{\beta}\right)}\right], \tag{52}$$

która jak widać jest statystyką ilorazu wiarygodności (51).

Wniosek: Zatem modele mogą być porównywane poprzez obliczenie różnic pomiędzy parami dewiancji tych modeli.

D1.20 Charakter kowarianta „wiek” - interakcja czy zaburzenie

Głównym wpływem interesującym nas w analizie ryzyka jest zmienna „obszar”. W formule (14) na ryzyko względne zmienna wiek okazała się nawet nieistotna. Jednak w wyprowadzeniu (14) nie braliśmy pod uwagę możliwości występowania zmiennej pobocznej „wiek” jako kowarianta w interakcji ze zmienną „obszar”. Przyjrzyjmy się więc bliżej charakterowi zmiennej „wiek” z punktu widzenia sposobu wprowadzenia jej do modelu regresji.

Punkt 1. Zmienna poboczna „wiek” może być wprowadzona do multiplikatywnego **członu interakcji** ze zmienną „obszar”. Rozważanie tej możliwości związane jest z odpowiedzią na pytanie o to czy zmienna „wiek” modyfikuje wpływ zmiennej „obszar”, to znaczy, czy wpływ zmiennej „obszar” mierzony ryzykiem względnym, różni się dla różnych wariantów wieku?

Punkt 2. Zmienna „wiek” może być wprowadzona do modelu tylko jako **zaburzenie**. Możliwość ta jest rozważana wtedy, gdy po analizie powyższego punktu, okazało się, że wprowadzenie zmiennej „wiek” do modelu w członie interakcji jest nieistotne statystycznie. W takiej sytuacji rozważamy czy zmienna „wiek” jest zaburzeniem, tzn. czy powinna znaleźć się w modelu w jakiejkolwiek formie po to, aby dać właściwe określenie jej wpływu na oszacowanie interesującego nas parametru, którym w rozważanym przykładzie jest ryzyko względne?

Jest różnica pomiędzy wprowadzeniem do modelu nowej zmiennej w postaci zaburzenia lub w postaci iloczynowego członu interakcji. **Nie wykonuje się testów statystycznych w przypadku, gdy zmienna ma wejść do modelu w postaci zaburzenia** [4].
Szczegółowe omówienie problemu rozróżnienia pomiędzy interakcją, czyli własnością modyfikacji wpływu głównego zmiennej typu „obszar” przez kowarianta będącego zmienną poboczną typu „wiek”, a problemem zaburzenia głównego wpływu zmiennej „obszar” przez zmienną poboczną „wiek”, można znaleźć w [4].

D1.21.1 Analiza interakcji obszaru i wieku. Model 2

Aby rozstrzygnąć kwestię zawartą w powyższym Punkcie 1, dotyczącą możliwości, że zmienna „wiek” jest kowariantem modyfikującym wpływ zmiennej „obszar”, rozszerzmy Model 1 , (25) (porównaj (9)), o człon interakcji, otrzymując:

Model 2: $$\ln r_{ij} = \alpha + \sum_{k=1}^{7} \alpha_k U_k + \beta M + \sum_{k=1}^{7} \delta_k \left(M U_k\right) \quad , i = 1, 2, ..., 8; \quad j = 0, 1 . \qquad (53)$$

Aby uniknąć osobliwości, tzn. idealnej współliniowości, możemy dodać człony interakcji tylko dla siedmiu (a nie ośmiu) zmiennych kierunkowych U_k .

Istotność interakcji „wieku” z „obszarem” możemy testować weryfikując hipotezę zerową:

$$H_0: \delta_1 = \delta_2 = \ldots = \delta_7 = 0 \ , \qquad (54)$$

z wykorzystaniem statystyki ilorazu wiarygodności (51). Ma ona przy prawdziwości hipotezy zerowej H_0 asymptotycznie rozkład χ^2 z 7 stopniami swobody, co jest liczbą nowych parametrów wprowadzonych do wyższego Modelu 2.

Statystyka testowa (51) pozwala więc na porównanie Modelu 1 (bez interakcji) z Modelem 2, który zawiera siedem iloczynowych członów interakcji MU_k .

D1.21.2 Program SAS dla Modelu 2

Ponieważ w Modelu 2 chcemy uwzględnić również interakcję „wieku" i „obszaru", zatem po wczytaniu danych takich samych jak w Punktcie 1.14.1, należy przy korzystaniu z procedury GENMOD (Punkt 1.14.1) zmienić linię *model* na uwzględniający człony interakcji MU_k, k=1,2,...,7, wczytując program:

```
proc genmod data = model;
model Y = M U1 U2 U3 U4 U5 U6 U7 U1M U2M U3M U4M U5M U6M U7M / covb
dist = poisson
link = log
offset = ln;
run;
quit;
```

D1.21.3 Raport z dopasowania Modelu 2

W wyniku analizy otrzymujemy następujący komputerowy raport SAS z dopasowywania Modelu 2. Jak to wynika z powyższych rozważań, raport ten dotyczy analizy z uwzględnieniem interakcji zmiennych „wiek" i „obszar".

```
                    System SAS
                The GENMOD Procedure

                Informacje o modelu

 Zbiór                          WORK.MODEL2
 Rozkład                            Poisson
 Funkcja wiążąca                        Log
 Zmienna zależna                          Y
 Zmienna przesunięcia                    ln

Liczba obserwacji wczytanych              16
Liczba obserwacji użytych                 16


  Informacje o poziomie klasyfikacji

 Klasa     Poziomy    Wartości

 A              8    1 2 3 4 5 6 7 8
```

System SAS
The GENMOD Procedure

Kryteria oceny zgodności

Kryterium	St. sw.	Wartość	Wartość/st. sw.
Dewiancja	0	**0.0000**	.
Skalowana dewia	0	0.0000	.
Chi-kwadrat Pearso	0	0.0000	.
Scaled Pearson X2	0	0.0000	.
Log. wiarogodn		7205.9610	

Algorytm osiągnął zbieżność.

System SAS
The GENMOD Procedure

Analiza ocen parametrów

Parametr kw..	St. sw.	Ocena	Błąd standardowy	95% granice przedziału ufności Walda		Chi-kwadrat	Pr > chi
Intercept	1	-5.3385	0.1581	-5.6484	-5.0286	1139.98	<.0001
M	1	0.5852	0.2010	0.1913	0.9790	8.48	0.0036
U1	1	-6.7207	1.0124	-8.7050	-4.7364	44.07	<.0001
U2	1	-3.6094	0.2958	-4.1891	-3.0296	148.89	<.0001
U3	1	-2.7347	0.2415	-3.2080	-2.2613	128.20	<.0001
U4	1	-1.8289	0.1977	-2.2164	-1.4414	85.58	<.0001
U5	1	-1.2232	0.1866	-1.5888	-0.8575	42.99	<.0001
U6	1	-0.7040	0.1808	-1.0584	-0.3496	15.16	<.0001
U7	1	-0.1504	0.1803	-0.5038	0.2030	0.70	0.4042
U1M	1	0.7521	1.1360	-1.4743	2.9786	0.44	**0.5079**
U2M	1	0.1075	0.3594	-0.5970	0.8120	0.09	**0.7649**
U3M	1	0.5605	0.2866	-0.0012	1.1221	3.83	**0.0505**
U4M	1	0.3599	0.2429	-0.1161	0.8360	2.20	**0.1384**
U5M	1	0.2067	0.2325	-0.2490	0.6623	0.79	**0.3740**
U6M	1	0.2620	0.2265	-0.1819	0.7059	1.34	**0.2474**
U7M	1	0.0490	0.2288	-0.3994	0.4973	0.05	**0.8305**
Skala	0	1.0000	0.0000	1.0000	1.0000		

UWAGA: The scale parameter was held fixed.

Z raportu SAS widać, że dewiancja dla Modelu 2 jest dokładnie równa zero:

$$D\left(\hat{\beta}\right)^{Model\,2} = 0 \ , \tag{55}$$

co oznacza, że model ten dopasowuje się do danych empirycznych idealnie. *Fakt ten jest spowodowany dopasowywaniem modelu z* 16 *parametrami do* $N = 16$ *elementowego zbioru danych.*

Jednak z raportu widać (pogrubienie na końcu linii `U1M do U7M`), że oszacowania parametrów interakcji $\delta_1, \delta_2, \ldots, \delta_7$ różnią się na poziomie istotności $\alpha = 0{,}05$ statystycznie nieistotnie od zera, co oznacza, że nie ma potrzeby aby wprowadzać interakcję. Sprawdźmy ten wniosek odwołując się do analizy z wykorzystaniem statystyki logarytmu ilorazu wiarygodności (51) dla Modelu 1 i Modelu 2.

D1.21.4 Testowanie braku dopasowania w Modelu 1 w porównaniu z Modelem 2

Rozważmy hipotezę zerową (54):

$$H_0: \delta_1 = \delta_2 = \ldots = \delta_7 = 0 \qquad (54')$$

mówiącą o nieistotności rozszerzenia Modelu 1 do Modelu 2, czyli statystycznej nieistotności interakcji.

Okazuje się, że w rozważanym przypadku test statystyczny weryfikujący hipotezę zerową (54), można by przeprowadzić zarówno wykorzystując statystykę ilorazu wiarygodności (co jest oczywiste), jak i dewiancję Modelu 1.
Istotnie, zauważmy po pierwsze, że obie te statystyki mają w przybliżeniu rozkład chi-kwadrat [4]. Po drugie, zauważmy, że dewiancja dla Modelu 1 otrzymana w raporcie w D**1.14.2** przyjęła w próbce wartość:

$$D\left(\hat{\beta}_{(r)}\right)^{Model1} = 8.195 \quad . \qquad (56)$$

Natomiast liczba stopni swobody dewiancji $D\left(\hat{\beta}_{(r)}\right)^{Model1}$ wynosi [1]:

$$d.f. = [\textit{liczba zmiennych } (Y_{ij})] - [\textit{liczba parametrów w } \text{Modelu 1}] = N - (r+1)$$
$$= 16 - 9 = 7. \qquad (57)$$

Statystyka $D\left(\hat{\beta}_{(r)}\right)^{Model1}$ ma więc w przybliżeniu rozkład chi-kwadrat (48) z *d.f.* = 7.

Z kolei statystyka testowa ilorazu wiarygodności (51) dla hipotezy zerowej (54) jest zgodnie z (52) otrzymana przez odjęcie dewiancji dla Modelu 2 (która jest równa zero) od dewiancji dla Modelu 1, tzn.:

$$-2\ln\left[\frac{P\left(\tilde{Y} \mid \hat{\beta}_{(r)}\right)}{P\left(\tilde{Y} \mid \hat{\beta}\right)}\right] = D\left(\hat{\beta}_{(r)}\right)^{Model1} - D\left(\hat{\beta}\right)^{Model2} = 8.195 - 0 = 8.195 \quad , \qquad (58)$$

zatem jej wartość w próbce jest równa $D\left(\hat{\beta}_{(r)}\right)^{Model1}$ jak w (56).

Również liczba stopni swobody statystyki ilorazu wiarygodności (51), równa [1]:

$$d.f. = [\textit{liczba parametrów w } \text{Modelu 2}] - [\textit{liczba parametrów w } \text{Modelu 1}]$$
$$= 16 - 9 = 7\,, \qquad (59)$$

wynosi tyle ile *d.f.* dewiancji Modelu 1, więc i ona ma w przybliżeniu rozkład chi-kwadrat (48) z *d.f.* = 7.

Zbierzmy informacje zawarte we wzorach (56) do (59). Wynika z nich, że skoro zarówno rozkład, jak i wartość liczbowa oraz liczba stopni swobody dewiancji Modelu 1, (56), oraz log ilorazu wiarygodności, (57), są takie same, zatem równoważnie można weryfikować hipotezę zerową (54) korzystając ze statystyki (57) bądź (56).

Przyjmijmy więc, w tym przypadku, dewiancję $D\left(\hat{\beta}_{(r)}\right)^{Model1}$ dla Modelu 1 jako statystykę testową hipotezy (54). Korzystając z (56) oraz (57) otrzymujemy, wykonując pomocnicze rachunki na przykład w arkuszu kalkulacyjnym Excel, że empiryczny poziom istotności wynosi:

$$p = \Pr\left(\chi_7^2 \geq D\left(\hat{\beta}_{(r)}\right)^{Model1} = 8{,}195\right) = 0.3157 \ . \qquad (60)$$

D1.21.4.1 Wniosek dla analizy interakcji zmiennych „obszar” i „wiek”

Zatem na żadnym poziomie istotności α mniejszym od jak widać dość dużego p = 0.3157, np. na poziomie $\alpha = 0{,}1$, nie mamy podstaw do odrzucenia hipotezy zerowej o statystycznej nieistotności rozszerzenia Modelu 1 do Modelu 2. Uznajemy więc, że *w* Modelu 1 *nie ma statystycznie istotnego braku dopasowania do danych empirycznych w porównaniu z* Modelem 2. Ponieważ Model 2 oraz model podstawowy dopasowują się do danych pomiarowych tak samo dobrze, zatem widzimy, że w Modelu 1 nie ma istotnego odchylenia obserwowanych wartości Y_{ij} od wartości przewidywanych $\hat{Y}_{ij}$ tym modelem.

Pozostawiamy więc prostszy Model 1 jako wystarczający do *przewidywania oczekiwanej ilości przypadków awarii* silnika, stwierdzając, że dodanie członów interakcji MU_k do Modelu 1 skomplikowałoby niepotrzebnie model, nie poprawiając w sposób statystycznie istotny dopasowania do danych empirycznych.

D1.21.5 Analiza „wieku” jako zaburzenia czynnika głównego

Rozważenie Punktu 2 w D1.20 polega na szukaniu odpowiedzi na pytanie o to, czy „wiek” jest kowariantem zaburzającym główny wpływ czynnika jakim jest „obszar”. Odpowiedz tą otrzymuje się wraz ze zbadaniem czy ryzyko względne $\hat{r}_{Wi} = e^{\hat{\beta}}$ albo równoważnie $\hat{\beta}$ zmienia się znacząco, jeśli zignorujemy zmienną „wiek”. Nie wprowadzenie „wieku” do analizy w Modelu 1 pozostawia tą zmienną poza kontrolą [4].

D1.21.5.1 Znacząca różnica ekspercka

Aby przeprowadzić potrzebną analizę należy więc pominąć wyrażenia dla „wieku”, tzn. składnik $\sum_{k=1}^{7} \alpha_k U_k$ z Modelu 1 i zobaczyć, czy otrzymane oszacowanie współczynnika przy *M* różni się będzie **znacząco** od wartości $\hat{\beta} = 0{,}8043$, (30), albo lepiej czy oszacowanie względnego ryzyka (bo to ono ostatecznie interesuje badacza) różni się znacząco od wartości $\hat{r}_{Wi}^{Model1} = e^{\hat{\beta}} = e^{0{,}8043} = 2{,}23513$. *Termin „znacząca różnica” nie odnosi się do testów statystycznych, ale do wiedzy ekspertów w dziedzinie*.

D1.21.5.2 Analiza SAS dla Modelu 3

Aby odpowiedzieć na pytanie o ile zmieni oszacowanie współczynnika β przy M, musimy dopasowywać do danych następujący model:

Model 3: $\ln r_{ij} = \alpha + \beta M$, i =1, 2,..., 8, j = 0, 1 (61)

Zadanie dla Modelu 3. Napisać program korzystający z procedury GENMOD dla Modelu 3, a następnie wykorzystując dane podane w Punkcie 1.14.1 uruchomić go, otrzymując poniższy raport SAS.

D1.21.5.3 Raport SAS dla Modelu 3

```
                    System SAS
                 The GENMOD Procedure

                 Informacje o modelu

 Zbiór                          WORK.MODEL3
 Rozkład                            Poisson
 Funkcja wiążąca                        Log
 Zmienna zależna                          Y
 Zmienna przesunięcia                    ln

Liczba obserwacji wczytanych              17
Liczba obserwacji użytych                 16
Braki danych                               1

  Informacje o poziomie klasyfikacji

 Klasa     Poziomy    Wartości

 A              8    1 2 3 4 5 6 7 8
```

Informacje o parametrach

Parametr	Efekt
Prm1	Intercept
Prm2	M

Kryteria oceny zgodności

Kryterium	St. sw.	Wartość	Wartość/st. sw.
Dewiancj	14	2569.7700	183.5550
Skalowana dewia	14	2569.7700	183.5550
Chi-kwadrat Pearso	14	3012.0987	215.1499
Scaled Pearson X2	14	3012.0987	215.1499
Log. wiarogodn		5921.0760	

Algorytm osiągnął zbieżność.

System SAS
The GENMOD Procedure

Analiza ocen parametrów

Parametr	St. sw.	Ocena	Błąd standardowy	95% granice przedziału ufności Walda		Chi-kwadrat	Pr > chi kw..
Intercept	1	-7.1273	0.0437	-7.2130	-7.0416	26567.6	<.0001
M	1	**0.7431**	0.0521	0.6410	0.8453	203.23	<.0001
Skala	0	1.0000	0.0000	1.0000	1.0000		

UWAGA: The scale parameter was held fixed.

D1.21.5.4 Analiza raportu SAS dla Modelu 3

Z powyższego raportu odczytujemy, że oszacowanie parametru β wynosi $\hat{\beta} = 0{,}7431$, skąd surowe oszacowanie (z powodu braku w analizie zmiennej „wiek") względnego ryzyka, wynosi:

$$\hat{r}_W^{Model3} = e^{\hat{\beta}} = e^{0{,}7431} = 2.1024 \ . \qquad (62)$$

Uwaga: Podkreślmy raz jeszcze, że w przeciwieństwie do różnicy istotnej statystycznie, wypowiedź o znaczącej różnicy, nie jest poparta żadnym statystycznym testem i nie należy testów takich wykonywać. O tym, czy różnica jest znacząca wypowiadają się specjaliści w branży.

Wniosek dotyczący zaburzenia: Porównując wartości Modelu 1 oraz Modelu 3 dla $\hat{\beta}$, które wynoszą odpowiednio 0,8043 oraz 0,7431 lub lepiej dla względnego ryzyka $\hat{r}_W = e^{\hat{\beta}}$, które wynoszą odpowiednio 2,2351 oraz 2,1024, uznajmy (chociaż nie jesteśmy specjalistami z branży samochodowej), że różnią się one znacząco i *zmienną poboczną „wiek" eksploatacji samochodu należy wprowadzić do modelu jako zaburzenie głównego wpływu zmiennej „obszar" eksploatacji samochodu.*

D1.21.5.5 Analiza rozszerzenia Modelu 3 do wyższego w hierarchii Modelu 1

Z porównania raportów dla Modelu 3 oraz Modelu 1 widać, że różnica dewiancji tych modeli wynosi: 2569.77 - 8.195 = 2561.58. Różnica dewiancji tych modeli (52), tzn. log ilorazu funkcji wiarygodności, ma w przybliżeniu rozkład chi-kwadrat. Przy różnicy 14-7=7 stopni swobody dewiancji tych modeli, wartość 2561.58 jest wysoce istotna statystycznie, wskazując na istotny brak dopasowania Modelu 3 w stosunku do Modelu 1.

Zadanie: Sformułować postać hipotezy zerowej mówiącej o nie występowaniu braku dopasowania do danych pomiarowych w Modelu 3 w porównaniu z Modelem 1. Wyznaczyć empiryczny poziom istotności dla przeprowadzanego testu tej hipotezy.

D1.22 Analiza regresji Poissona w SAS dla modelu z przesunięciem

Dla skompletowania analizy dla wszystkich modeli ze zbioru modeli hierarchicznych rozważymy jeszcze model tylko z wyrazem wolnym, czyli taki w którym występuje brak zależności modelowej od zmiennych objaśniających. Model ten ma postać:

Model 0: $\ln r_{ij} = \alpha$, i =1, 2,..., 8; j = 0, 1 . (63)

D1.22.1 Dane i program SAS dla Modelu 0

Aby przeprowadzić analizę z użyciem procedury GENMOD została w danych podanych w D1.14.1 wprowadzona dodatkowa zmienna O, przyjmująca zawsze wartość zero.

Ponieważ w Modelu 0 występuje brak zależności modelowej od zmiennych objaśniających, w związku z tym modyfikujemy następująco wiersz *model* poleceń w procedurze GENMOD:

model Y = O / cov
lub
model Y = O / pred covb

D1.22.2 Raport SAS dla Modelu 0

Po wczytaniu danych zawartych w D1.14.1 oraz uruchomieniu programu procedury GENMOD, otrzymujemy poniższy raport.

System SAS
The GENMOD Procedure

Informacje o modelu

Zbiór	WORK.MODEL0
Rozkład	Poisson
Funkcja wiążąca	Log
Zmienna zależna	Y
Zmienna przesunięcia	ln
Liczba obserwacji wczytanych	16
Liczba obserwacji użytych	16

Informacje o poziomie klasyfikacji

Klasa	Poziomy	Wartości
A	8	1 2 3 4 5 6 7 8

Informacje o parametrach

Parametr	Efekt
Prm1	Intercept
Prm2	O

Kryteria oceny zgodności

Kryterium	St. sw.	Wartość	Wartość/st. sw.
Dewiancja	15	**2790.3403**	186.0227
Skalowana dewia	15	2790.3403	186.0227
Chi-kwadrat Pearso	15	3480.1347	232.0090
Scaled Pearson X2	15	3480.1347	232.0090
Log. wiarogodn		5810.7909	

Algorytm osiągnął zbieżność.

Analiza ocen parametrów

Parametr	St. sw.	Ocena	Błąd standardowy	95% granice przedziału ufności Walda		Chi-kwadrat	Pr > chi kw..
Intercept	1	-6.6669	0.0238	-6.7135	-6.6202	78449.1	<.0001
I	0	0.0000	0.0000	0.0000	0.0000	.	.
Skala	0	1.0000	0.0000	1.0000	1.0000		

UWAGA: The scale parameter was held fixed.

D1.22.3 Wynik analizy dla Modelu 0

Dewiancja dla Modelu 0, $\ln r_{ij} = \alpha$, wynosi 2790.3403. Jak można się było spodziewać, model posiadający tylko przesunięcie i bez zależności od zmiennych objaśniających wykazuje

istotny brak dopasowania w stosunku do Modelu 1, $\ln r_{ij} = \alpha + \sum_{k=1}^{7} \alpha_k U_k + \beta M$, co przejawia się gwałtownym wzrostem dewiancji Modelu 0, (63), w stosunku do dewiancji Modelu 1, (25).

Zadanie: Sformułować postać hipotezy zerowej mówiącej o nie występowaniu braku dopasowania do danych pomiarowych w Modelu 0 w porównaniu z Modelem 1 . Wyznaczyć empiryczny poziom istotności dla przeprowadzanego testu tej hipotezy sprawdzając powyższy wynik analizy dla Modelu 0.

Zadanie: Pokazać, że różnica dewiancji Modelu 0 , (63), oraz Modelu 3, (61), jest również statystycznie istotna, znajdując wartość odpowiedniego empirycznego poziomu istotności.

D1.23 Podsumowanie analizy regresji doboru modelu Poissona

Poniższa Tabela 2 podsumowuje przeprowadzoną analizę regresji Poissona dla przykładu zależności liczby awarii silnika w klasie modeli hierarchicznych z uwzględnieniem „obszaru" jako czynnika głównego wpływu, a zmiennej „wiek" jako zmiennej pobocznej.

Tabela 2

Tabela ANOVA dla przykładu awarii silnika ($N = 16$).

	Model dla $\ln r_{ij}$	Liczba parametrów	$D(\beta)$	*d.f.*	Istotna statystycznie różnica w $D(\beta)$
Model 0	α	1	2790,34	15	↕ Istotna $P \approx 0$
Model 3	$\alpha + \beta M$	2	2569,77	14	↕ Istotna $P \approx 0$
Model 1	$\alpha + \sum_{k=1}^{7} \alpha_k U_k + \beta M$	9	8,2	7	↕ Nieistotna $p = 0{,}32$
Model 2	$\alpha + \sum_{k=1}^{7} \alpha_k U_k + \beta M + + \sum_{k=1}^{7} \delta_k \, MU_k$	16	0	0	
Model podstawowy	μ_j	16	0	0	

D1.23.1 Wniosek z analizy

Z przeprowadzonej analizy widać, że dane zawierają wskazanie, że spośród rozważanego zbioru modeli hierarchicznych należałoby wybrać Model 1 jako ten, który nie ma statystycznie istotnego braku dopasowania do danych pomiarowych, a jednocześnie ma prostszą strukturę (9 parametrów) niż model podstawowy lub Model 2 z interakcją (16 parametrów).

Uwaga: Wykroczenie poza klasę modeli hierarchicznych i potraktowanie „wieku” jako zmiennej typu ciągłego mogłoby doprowadzić do wyselekcjonowania modelu z mniejszą liczbą parametrów niż Model 1 [4].

Uzupełnienie: Polecenia języka 4GL procedury GENMOD dla rozważanego przykładu

Poniżej podane zostały podstawowe komendy programów napisanych w języku 4GL dla celów przeprowadzenia analizy regresji Poissona, w tym rozważanego powyżej przykładu.

data przykład1 wskazuje nazwę zbioru z danymi;
input wskazuje zmienne, które mają być wczytane do modelu;
ln wskazuje zewnętrzną zmienną funkcyjną (tutaj logarytm);
datalines wskazuje, że poniżej będą się znajdowały linie danych;
run wskazuje na koniec linii danych;

proc oznacza początek odpowiedniej procedury (w Dodatku: genmod);
model wskazuje zmienne użyte w modelu;
pred wskazuje na konieczność wyliczenia wartości prognozowanych;
ref wskazuje referencyjną populację (tzn. linię, w której wszystkie zmienne kierunkowe dla przyjętego systemu kodowania oraz ich interakcje mają wartość 0);
covb wskazuje na wyliczenie macierzy kowariancji estymatorów;
dist informuje o użyciu określonego rozkładu;
link informuje o użyciu wskazanej funkcji linku (w Dodatku: logarytmicznej);
offset wskazuje zmienną, znajdującą się poza modelem, w której przechowywana jest funkcja linkująca;
run informuje o uruchomieniu procedury liczącej;
quit powoduje wyjście z programu i wyświetlenie wydruku.

Opis zmiennych występujących w zbiorze danych w D1.14.1.

Zmienna A jest zmienną jakościową z wariantem wieku serwisowanych samochodów;

Y oznacza zmienną objaśnianą ilości występujących przypadków (zmienna o rozkładzie Poissona);

N oznacza liczebność badanych populacji;

M jest zmienną kierunkową wskazującą na obszar;

U1, U2, U3, U4, U5, U6, U7 są zmiennymi kierunkowymi, wskazującymi na odpowiednią przynależność do klasy wiekowej;

U1M, U2M, U3M, U4M, U5M, U6M, U7M to interakcje zmiennych kierunkowych „wiek” U1, U2, U3, U4, U5, U6, U7 oraz „obszar” M;

O jest zmienną sztucznie wprowadzoną dla celu analizy Modelu 0, która nie jest zmienną objaśniającą.

Zakończenie

Przedmiotem Dodateku do Rozdziału 1 skryptu [1] jest przećwiczenie zastosowania metody największej wiarygodności (MNW) w problemach estymacyjnych analizy regresji Poissona. Rozważania zostały poparte przykładami przeliczonymi z wykorzystaniem systemu analiz statystycznych SAS.

Omówiono sposób konstrukcji funkcji wiarygodności wykorzystywany dla celów budowy estymatorów parametrów modelu oraz wynikające z tej metody procedury wnioskowania statystycznego. Procedury dla testowania hipotez i konstruowania przedziałów ufności wykorzystują nie tylko zmaksymalizowane wartości funkcji wiarygodności, ale również oszacowane macierze kowariancji wyznaczane w ramach szerzej rozumianej metody największej wiarygodności odwołującej się do tzw. informacji Fishera zawartej w próbie.

Teoretyczne podstawy MNW wraz ze znaczeniem informacji Fishera dla (estymacji) macierzy kowariancji estymatorów parametrów modelu znajdują się w literaturze zacytowanej w Dodatku.

W omówionych przykładach zmienna losowa objaśniana zawsze była liczbą zliczeń przypadków interesującego nas zdarzenia. Dlatego przy spełnieniu warunku małej liczby defektów w stosunku do wszystkich obserwacji w rozważanych podgrupach próbek pobranych z dwóch porównywanych populacji, wykorzystywana postać funkcji wiarygodności odwoływała się do zmiennej mającej rozkład Poissona. W praktyce, dla typowego modelu regresji Poissona naturalną miarą estymowanego efektu jest tempo awarii (tzn. ryzyko) oraz ryzyko względne, związane z określonym, interesującym nas czynnikiem, którego warianty kontrastują badane populacje.

W Dodatku przedstawiono metodę selekcji modelu z wykorzystaniem statystyki ilorazu wiarygodności oraz zastosowanie statystyki dewiancji, która jest rodzajem statystyki ilorazu wiarygodności, opisującej dobroć dopasowania badanego modelu względem modelu podstawowego. Ponieważ różnica w statystyce dewiancji, otrzymana dla dwóch porównywanych modeli, jest równa statystce logarytmu ilorazu funkcji wiarygodności dla tych modeli, więc testy hipotez o braku dopasowania w modelach niższych w hierarchii, mogą być przeprowadzony z wykorzystaniem różnicy statystyk dewiancji, które pojawiają się raportach SAS.

Zastosowanie MNW w analizie regresji Poissona ma kluczowe znaczenie ze względu na możliwość selekcji modelu, który nie tylko ma estymatory posiadające (asymptotycznie)

optymalne własności [2], ale jak na to zwrócono uwagę w analizowanych przykładach, nie wykazuje również statystycznie istotnie gorszego dopasowania do danych empirycznych niż model podstawowy, posiadając przy tym najmniejszą możliwą liczbę parametrów.

Typowy model regresji Poissona, użyty w przykładach, wyraża w postaci logarytmicznej tempo porażki jako liniową funkcję zbioru czynników. Nie mniej estymacja MNW jest szczególnie przydatna w estymacji współczynników regresji w modelach nieliniowych, takich jak model regresji logistycznej czy model regresji Poissona. Ponieważ układ równań wiarygodności nie prowadzi wtedy do liniowych równań algebraicznych na estymatory tych parametrów, dlatego procedury estymacji dla takich modeli wymagają programu komputerowego, stosującego algorytmy z wielokrotnymi iteracjami estymatorów parametrów modelu. Taki pakiet numerycznych procedur komputerowych jest zawarty w systemie SAS.
Podstawową procedurą SAS stosowaną w analizie regresji Poissona w sytuacji, gdy logarytm ryzyka jest liniową kombinacją czynników, jest procedura GENMOD. W bardziej skomplikowanych nieliniowych modelach regresji Poissona, gdy logarytmu ryzyka nie da się przedstawić w postaci liniowej kombinacji czynników, właściwą procedurą, którą można wykorzystać jest procedura NLMIXED [4].

Literatura

[1] J. Syska, „Metoda największej wiarygodności i informacja Fisher'a w fizyce i ekonofizyce", skrypt dla studentów kierunku Ekonfizyka, Instytut Fizyki, Uniwersytet Śląski, (2011) .

[2] R. Nowak, „Statystyka dla fizyków", Wydawnictwo Naukowe PWN, Warszawa, (2002).

[3] S. Amari, H. Nagaoka, *Methods of information geometry, transletions of Mathematical monographs*, Vol.191, Oxford Univ. Press, (2000).

[4] D.G. Kleinbaum, L.L. Kupper, K.E. Muller, A. Nizam, "Applied Regression Analysis and Multivariable Methods", Duxbury Press, (1998).

[5] W. Krysicki, J.Bartos, W. Dyczka, K. Królikowska, M. Wasilewski, „Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach", „Część II. Statystyka matematyczna", Wydawnictwo Naukowe PWN, Warszawa, (1995).

[6] Y. Pawitan, "In all likelihood, Statistical Modeling and inference using likelihood", Oxford, (2001).

[7] J. Jakubowski, R. Sztencel, *Wstep do teorii prawdopodobienstwa*, wydanie 2, Script, Warszawa, (2001).

[8] E. Frątczak, M. Pęczkowski, K. Sienkiewicz, K. Skaskiewicz, „Statystyka od podstaw z systemem SAS", Szkoła Główna Handlowa, Warszawa 2001.

[9] M.Maliński, "Statystyka matematyczna wspomagana komputerowo", Wydawnictwo Politechniki Śląskiej, Gliwice (2000).