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 trzecia)

Spis treści

Wstęp

Dodatek ten jest uzupełnieniem do Rozdziału pierwszego skryptu [1] „Metoda największej wiarygodności i informacja Fisher'a w fizyce i ekonofizyce". Aby dodatek tworzył niezależną całość, powtórzono w nim Rozdział 1 skryptu [1]. 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].

Wprowadzenie do metody największej wiarygodności

Z powodu możliwości zastosowania *metody największej wiarygodności* (MNW) do rozwiązania wielu, bardzo różnych problemów estymacyjnych, stała się ona obecnie zarówno metodą podstawową jak również punktem wyjścia dla różnych metod analizy statystycznej. Jej wszechstronność związana jest, po pierwsze z możliwością przeprowadzenia analizy statystycznej dla małej próbki, opisu zjawisk nieliniowych oraz zastosowania zmiennych losowych posiadających zasadniczo dowolny *rozkład prawdopodobieństwa* [2], oraz po drugie, szczególnymi własnościami otrzymywanych przez nią estymatorów, które okazują się być zgodne, asymptotycznie nieobciążone, efektywne oraz dostateczne [2]. MNW zasadza się na intuicyjnie jasnym postulacie przyjęcia za prawdziwe takich wartości parametrów rozkładu prawdopodobieństwa zmiennej losowej, które maksymalizują funkcję wiarygodności realizacji konkretnej próbki.

W1.1 Podstawowe pojęcia MNW

Rozważmy zmienną losową Y [2], która przyjmuje wartości y zgodnie z rozkładem prawdopodobieństwa $p(y|\theta)$, gdzie $\theta = (\vartheta_1, \vartheta_2, ..., \vartheta_k)^T \equiv (\vartheta_s)_{s=1}^k$, jest zbiorem k parametrów tego rozkładu (T oznacza transpozycję). Zbiór wszystkich możliwych wartości y zmiennej Y oznaczmy przez Y.

Gdy $k > 1$ wtedy θ nazywamy parametrem *wektorowym*. W szczególnym przypadku $k = 1$ mamy $\theta = \vartheta$. Mówimy wtedy, że parametr θ jest parametrem *skalarnym*.

Pojęcie próby i próbki: Rozważmy *zbiór danych* $y_1, y_2, ..., y_N$ otrzymanych w N obserwacjach zmiennej losowej Y.

Każda z danych y_n, $n = 1,2,...,N$, jest generowana z rozkładu $p_n(y_n|\theta_n)$ zmiennej losowej Y w populacji, którą charakteryzuje wartość parametru wektorowego $\theta_n = (\vartheta_1, \vartheta_2, ..., \vartheta_k)_n^T \equiv ((\vartheta_s)_{s=1}^k)_n$, $n = 1,2,...,N$. Stąd zmienną Y w n-tej populacji oznaczymy Y_n. Zbiór zmiennych losowych $\tilde{Y} = (Y_1, Y_2, ..., Y_N) \equiv (Y_n)_{n=1}^N$ nazywamy N-wymiarową *próbą*.

Konkretną realizację $y = (y_1, y_2, ..., y_N) \equiv (y_n)_{n=1}^N$ próby $\tilde{Y}$ nazywamy *próbką*. Zbiór wszystkich możliwych realizacji y próby $\tilde{Y}$ tworzy przestrzeń próby (układu) oznaczaną jako $\mathbf{B}$.

Określenie: Ze względu na to, że *n* jest indeksem konkretnego punktu pomiarowego próby, rozkład $p_n(y_n|\theta_n)$ będziemy nazywali rozkładem „punktowym” (czego nie należy mylić z np. rozkładem dyskretnym).

Określenie funkcji wiarygodności: Centralnym pojęciem MNW jest *funkcja wiarygodności* $L(y;\Theta)$ (pojawienia się) próbki $y = (y_n)_{n=1}^N$, nazywana też *wiarygodnością próbki*. Jest ona funkcją parametru Θ.

Przez wzgląd na zapis stosowany w fizyce, będziemy stosowali oznaczenie $P(y|\Theta) \equiv L(y;\Theta)$, które podkreśla, że formalnie *funkcja wiarygodności jest łącznym rozkładem prawdopodobieństwa* [1] pojawienia się realizacji $y \equiv (y_n)_{n=1}^N$ próby $\tilde{Y} \equiv (Y_n)_{n=1}^N$, to znaczy:

$$P(\Theta) \equiv P(y|\Theta) = \prod_{n=1}^{N} p_n(y_n|\theta_n). \qquad \text{(W1)}$$

Zwrócenie uwagi w (W1) na występowanie y w argumencie funkcji wiarygodności oznacza, że może być ona rozumiana jako statystyka $P(\tilde{Y} \mid \Theta)$. Z kolei skrócone oznaczenie $P(\Theta)$ podkreśla, że centralną sprawą w MNW jest fakt, że funkcja wiarygodności jest funkcją nieznanych parametrów:

$$\Theta = (\theta_1, \theta_2, ..., \theta_N)^T \equiv (\theta_n)_{n=1}^N \quad \text{przy czym} \quad \theta_n = (\vartheta_{1n}, \vartheta_{2n}, ..., \vartheta_{kn})^T \equiv ((\vartheta_s)_{s=1}^k)_n \,, \qquad \text{(W2)}$$

gdzie θ_n jest wektorowym parametrem populacji określonej przez indeks próby n. W toku analizy chcemy oszacować wektorowy parametr Θ.

Zbiór wartości parametrów $\Theta = (\theta_n)_{n=1}^N$ tworzy współrzędne rozkładu prawdopodobieństwa rozumianego jako punkt w $d = k \times N$ - wymiarowej (podprzestrzeni) przestrzeni statystycznej S [1].

Uwaga o postaci rozkładów punktowych: Tak jak w [1], zakładamy, że "*punktowe*" rozkłady $p_n(y_n \mid \theta_n)$ dla poszczególnych pomiarów n w N elementowej próbie są *niezależne*[2]. W ogólności [1], rozkłady punktowe $p_n(y_n \mid \theta_n)$ zmiennych Y_n chociaż są *tego samego typu*, jednak nie spełniają warunku $p_n(y_n \mid \theta_n) = p(y \mid \theta)$, charakterystycznego dla próby prostej. Taka ogólna sytuacja ma np. miejsce w analizie regresji Poissona (Rozdział D).

Pojęcie estymatora parametru: Załóżmy, że dane $y = (y_n)_{n=1}^N$ są generowane losowo z punktowych rozkładów prawdopodobieństwa $p_n(y_n \mid \theta_n)$, $n = 1,2,...,N$, które chociaż nie są znane, to jednak założono o nich, że dla każdego n należą do określonej, tej samej klasy modeli. Zatem funkcja wiarygodności (W1) należy do określonej, $d = k \times N$ - wymiarowej, przestrzeni statystycznej S.

Celem analizy jest oszacowanie nieznanych parametrów Θ, (W2), poprzez funkcję:

$$\hat{\Theta} \equiv \hat{\Theta}(\tilde{Y}) = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_N)^T \equiv (\hat{\theta}_n)_{n=1}^N \quad \text{gdzie} \quad \hat{\theta}_n = (\hat{\vartheta}_{1n}, \hat{\vartheta}_{2n}, ..., \hat{\vartheta}_{kn})^T \equiv ((\hat{\vartheta}_s)_{s=1}^k)_n \,, \quad \text{(W4)}$$

mającą $d = k \times N$ składowych.

---

[2] W przypadku analizy jednej zmiennej losowej Y, rozkłady te obok niezależności spełniają dodatkowo warunek:

$$p_n(y_n \mid \theta_n) = p(y \mid \theta), \qquad \text{(W3)}$$

co oznacza, że próba jest *prosta*.

Każda z funkcji $\hat{\vartheta}_{kn} \equiv \hat{\vartheta}_{kn}(\tilde{Y})$ jako funkcja próby jest *statystyką*, którą przez wzgląd na to, że służy do oszacowywania wartości parametru ϑ_{kn} nazywamy estymatorem tego parametru. *Estymator parametru nie może zależeć od parametru, który oszacowuje*[3].

Podsumowując, odwzorowanie:

$$\hat{\Theta} : B \to \mathbf{R}^d \,, \tag{W5}$$

gdzie B jest przestrzenią próby, jest estymatorem parametru (wektorowego) Θ.

Równania wiarygodności: Będąc funkcją $\Theta = (\theta_n)_{n=1}^N$, funkcja wiarygodności służy do konstrukcji estymatorów $\hat{\Theta} = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_N)^T \equiv (\hat{\theta}_n)_{n=1}^N$ parametrów $\Theta \equiv (\theta_n)_{n=1}^N$. Procedura polega na wyborze takich $(\hat{\theta}_n)_{n=1}^N$, dla których funkcja wiarygodności przyjmuje maksymalną wartość, skąd statystyki te nazywamy estymatorami MNW.

Zatem, wprowadzony przez Fishera, warunek konieczny otrzymania estymatorów $\hat{\Theta}$ MNW sprowadza się do znalezienia rozwiązania układu $d = k \times N$ tzw. *równań wiarygodności* [1]:

$$S(\Theta)_{|\Theta=\hat{\Theta}} \equiv \frac{\partial}{\partial \Theta} \ln P(y|\,\Theta)_{|\Theta=\hat{\Theta}} = 0 \,, \tag{W6}$$

gdzie zagadnienie maksymalizacji funkcji wiarygodności $P(y|\,\Theta)$ sprowadzono do (na ogół) analitycznie równoważnego mu problemu maksymalizacji jej logarytmu $\ln P(y|\,\Theta)$.

Określenie funkcji wynikowej: Funkcję $S(\Theta)$ będącą gradientem logarytmu funkcji wiarygodności:

$$S(\Theta) \equiv \frac{\partial}{\partial \Theta} \ln P(y|\,\Theta) = \begin{pmatrix} \dfrac{\partial \ln P(y\,|\,\Theta)}{\partial \theta_1} \\ \vdots \\ \dfrac{\partial \ln P(y\,|\,\Theta)}{\partial \theta_N} \end{pmatrix} \quad \text{g}dzie \quad \frac{\partial \ln P(y\,|\,\Theta)}{\partial \theta_n} = \begin{pmatrix} \dfrac{\partial \ln P(y\,|\,\Theta)}{\partial \vartheta_{1n}} \\ \vdots \\ \dfrac{\partial \ln P(y\,|\,\Theta)}{\partial \vartheta_{kn}} \end{pmatrix}, \tag{W7}$$

nazywamy *funkcją wynikową*.

Po otrzymaniu (wektora) estymatorów $\hat{\Theta}$, *zmaksymalizowaną* wartość funkcji wiarygodności definiujemy jako numeryczną wartość funkcji wiarygodności powstałą przez podstawienie do $P(y|\,\Theta)$ wartości oszacowanej $\hat{\Theta}$ w miejsce parametru Θ.

[3] Natomiast rozkład estymatora oszacowywanego parametru, zależy od tego parametru.

Przykład: Rozważmy problem estymacji skalarnego parametru, tzn. $\Theta = \theta$ (tzn. $k = 1$ oraz $N = 1$), dla zmiennej losowej Y opisanej rozkładem dwumianowym (Bernoulliego):

$$P(y \mid \theta) = \binom{m}{y} \theta^{y} (1-\theta)^{m-y} . \tag{W8}$$

Estymacji parametru θ dokonamy na podstawie *pojedynczej* obserwacji (długość próby $N = 1$) zmiennej Y, której iloraz *Y*/*m* nazywamy *częstością*. Parametr m charakteryzuje rozkład zmiennej Bernoulliego Y (i nie ma związku z długością N próby).

Zatem ponieważ $y \equiv (y_1)$, więc $P(y \mid \theta)$ jest funkcją wiarygodności dla $N = 1$ wymiarowej próby. Jej logarytm wynosi:

$$\ln P(y \mid \theta) = \ln \binom{m}{y} + y \ln \theta + (m-y) \ln (1-\theta). \tag{W9}$$

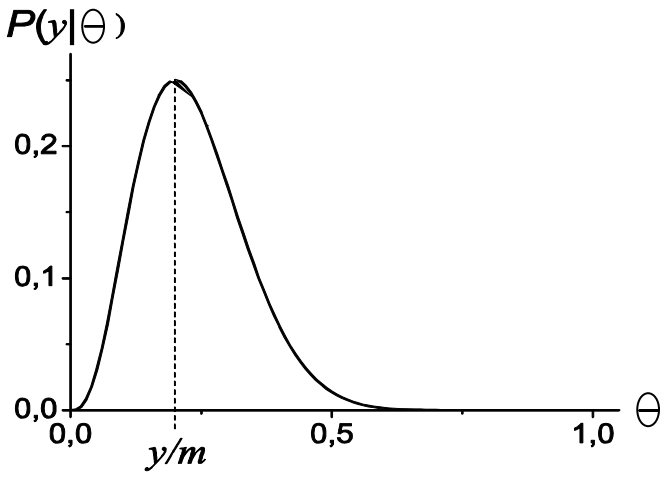
W rozważanym przypadku otrzymujemy jedno równanie wiarygodności (W6):

$$S(\theta) = \frac{1}{\theta} y - \frac{1}{1-\theta} (m-y) \Big|_{\theta=\hat{\theta}} = 0 \tag{W10}$$

a jego rozwiązanie daje estymator MNW parametru θ rozkładu dwumianowego, równy:

$$\hat{\theta} = \frac{y}{m} \tag{W11}$$

Ilustracją powyższej procedury znajdowania wartości estymatora parametru θ jest Rysunek 1.1 (gdzie przyjęto $m = 5$). Na skutek pomiaru zaobserwowano wartość Y równą $y = 1$.



Rysunek 1.1: Graficzna ilustracja metody największej wiarygodności dla $P(y \mid \theta)$ określonego wzorem (W8) dla rozkładu dwumianowego. Przyjęto wartość parametru $m = 5$. W pomiarze zaobserwowano wartość $Y = y = 1$.

Maksimum $P(y \mid \theta)$ przypada na wartość θ równą punktowemu oszacowaniu $\hat{\theta} = y/m = 1/5$ tego parametru. Maksymalizowana wartość funkcji wiarygodności wynosi $P(y \mid \hat{\theta})$.

W1.2 Wnioskowanie w MNW

Z powyższych rozważań wynika, że konstrukcja punktowego oszacowania parametru w MNW oparta jest o postulat maksymalizacji funkcji wiarygodności przedstawiony powyżej. Jest on wstępem do statystycznej procedury wnioskowania. Kolejnym krokiem jest konstrukcja przedziału wiarygodności. Jest on odpowiednikiem przedziału ufności, otrzymywanego w częstotliwościowym podejściu statystyki klasycznej do procedury estymacyjnej. Do jego konstrukcji niezbędna jest znajomość rozkładu prawdopodobieństwa estymatora parametru, co (dzięki ”porządnym”' granicznym własnościom stosowanych estymatorów) jest możliwe niejednokrotnie jedynie asymptotycznie, tzn. dla wielkości próby dążącej do nieskończoności. Znajomość rozkładu estymatora jest też niezbędna we wnioskowaniu statystycznym odnoszącym się do weryfikacji hipotez.

W sytuacji, gdy nie dysponujemy wystarczającą ilością danych, potrzebnych do przeprowadzenia skutecznego częstotliwościowego wnioskowania, Fisher [6] zaproponował do określenia niepewności dotyczącej parametru Θ wykorzystanie maksymalizowanej wartość funkcji wiarygodności.

Przedział wiarygodności jest zdefiniowany jako zbiór wartości parametru Θ, dla których funkcja wiarygodności osiąga (umownie) wystarczająco wysoką wartość, tzn.:

$$\left\{\Theta, \frac{P(y \mid \Theta)}{P(y \mid \hat{\Theta})} > c\right\}, \tag{W12}$$

dla pewnego *parametru obcięcia* c, nazywanego *poziomem wiarygodności*.

Iloraz wiarygodności:

$$\frac{P(y \mid \Theta)}{P(y \mid \hat{\Theta})} \tag{W13}$$

reprezentuje pewien typ unormowanej wiarygodności i jako taki jest wielkością skalarną. Jednak z powodu niejasnego znaczenia określonej wartości parametru obcięcia c pojęcie to wydaje się być na pierwszy rzut oka za słabe, aby dostarczyć taką precyzję wypowiedzi jaką daje analiza częstotliwościowa.

Istotnie, wartość c nie odnosi się do żadnej wielkości obserwowanej, tzn. na przykład 1%-we ($c = 0{,}01$) obcięcie nie ma ścisłego probabilistycznego znaczenia. Inaczej ma się sprawa dla częstotliwościowych przedziałów ufności. W tym przypadku wartość współczynnika $\alpha = 0{,}01$ oznacza, że gdybyśmy rozważyli realizację przedziału ufności na poziomie ufności $1-\alpha = 0{,}99$, to przy pobraniu nieskończonej (w praktyce wystarczająco dużej) liczby próbek, 99% wszystkich wyznaczonych przedziałów ufności pokryłoby prawdziwą (teoretyczną) wartość parametru Θ w populacji generalnej (składającej się z N podpopulacji). Pomimo tej słabości MNW, rozbudowanie analizy stosunku wiarygodności okazuje się być istotne we wnioskowaniu statystycznym analizy doboru modeli i to aż po konstrukcję równań teorii pola [1].

W1.2.1 Wiarygodnościowy przedział ufności

Przykład rozkładu normalnego z jednym estymowanym parametrem: Istnieje przypadek pozwalający na prostą *interpretację przedziału wiarygodnościowego jako przedziału ufności*. Dotyczy on zmiennej Y posiadającej rozkład Gaussa oraz sytuacji gdy (dla próby prostej) interesuje nas estymacja skalarnego parametru θ będącego wartością oczekiwaną $E(Y)$ zmiennej Y. Przypadek ten omówimy poniżej. W ogólności, przedział wiarygodności posiadający określony poziom ufności jest nazywany przedziałem ufności.

Częstotliwościowe wnioskowanie o nieznanym parametrze θ wymaga określenia rozkładu jego estymatora, co jest zazwyczaj możliwe jedynie granicznie [6]. Podobnie w MNW, o ile to możliwe, korzystamy przy dużych próbkach z twierdzeń granicznych dotyczących rozkładu ilorazu wiarygodności [6]. W przypadku rozkładu normalnego i parametru skalarnego okazuje się, że możliwa jest konstrukcja skończenie wymiarowa.

Niech więc zmienna Y ma rozkład normalny $N(\theta,\sigma^2)$:

$$p(y \mid \theta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(y-\theta)^2}{2\sigma^2}\right). \tag{W14}$$

Rozważmy próbkę $y \equiv (y_1,\ldots,y_N)$, która jest realizacją próby prostej $\tilde{Y}$ i załóżmy, że *wariancja σ^2 jest znana*. Logarytm funkcji wiarygodności dla $N(\theta,\sigma^2)$ ma postać:

$$\ln P(y \mid \theta) = -\frac{N}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n-\theta)^2 , \tag{W15}$$

gdzie ze względu na próbę prostą, w argumencie funkcji wiarygodności wpisano w miejsce $\Theta \equiv (\theta)_{n=1}^{N}$ parametr θ, jedyny który podlega estymacji.

Z postaci funkcji wiarygodności (W15) oraz związku $\sum_{n=1}^{N}\left(y_n - \hat{\theta}\right)^2 = \sum_{n=1}^{N}\left((y_n - \theta) + (\theta - \hat{\theta})\right)^2$, otrzymujemy[4]:

$$\ln \frac{P(y \mid \theta)}{P(y \mid \hat{\theta})} = -\frac{N}{2\sigma^2}\left(\hat{\theta} - \theta\right)^2 , \tag{W16}$$

gdzie $\hat{\theta} = y = \frac{1}{N}\sum_{n=1}^{N} y_n$ jest estymatorem MNW parametru θ.

Statystyka Wilka: Widać, że po prawej stronie (W16) otrzymaliśmy wyrażenie kwadratowe. Ponieważ $\overline{Y}$ jest nieobciążonym estymatorem parametru θ, co oznacza, że wartość oczekiwana $E(\overline{Y}) = \theta$, zatem (dla rozkładu $Y \sim N\left(\theta, \sigma^2\right)$) średnia arytmetyczna $\overline{Y}$ ma rozkład normalny $N\left(\theta, \frac{\sigma^2}{N}\right)$. Z normalności rozkładu $\overline{Y}$ wynika, że tzw. *statystyka ilorazu wiarygodności Wilka*:

$$W \equiv 2\ln \frac{P\left(\tilde{Y} \mid \hat{\theta}\right)}{P\left(\tilde{Y} \mid \theta\right)} \sim \chi_1^2 , \tag{W20}$$

ma rozkład χ^2, w tym przypadku z jednym stopniem swobody [6].

---

[4] **Postać estymatora parametru skalarnego θ rozkładu $N\left(\theta, \sigma^2\right)$**: Korzystając z równania wiarygodności (W6) dla przypadku skalarnego parametru θ, otrzymujemy:

$$S(\theta)_{|_{\theta=\hat{\theta}}} \equiv \frac{\partial}{\partial \theta} \ln P(y|\theta)_{|_{\theta=\hat{\theta}}} = 0 , \tag{W17}$$

skąd dla log funkcji wiarygodności (W15), otrzymujemy:

$$\hat{\theta} = y = \frac{1}{N}\sum_{n=1}^{N} y_n . \tag{W18}$$

Zatem estymatorem parametru θ jest średnia arytmetyczna:

$$\hat{\theta} = \overline{Y} = \frac{1}{N}\sum_{n=1}^{N} Y_n . \tag{W19}$$

Estymator i jego realizowaną wartość będziemy oznaczali tak samo, tzn. $\hat{\theta}$ dla przypadku skalarnego i $\hat{\Theta}$ dla wektorowego.

Wyskalowanie statystyki Wilka w przypadku normalnym: Wykorzystując (W20) możemy wykonać wyskalowanie wiarygodności oparte o możliwość powiązania przedziału wiarygodności z jego częstotliwościowym odpowiednikiem.

Mianowicie z (W20) otrzymujemy, że dla ustalonego (chociaż nieznanego) parametru θ prawdopodobieństwo, że iloraz wiarygodności znajduje się w wyznaczonym dla parametru obcięcia c, wiarygodnościowym przedziale ufności, wynosi:

$$P\left(\frac{P\left(\tilde{Y}\mid\theta\right)}{P\left(\tilde{Y}\mid\hat{\theta}\right)}>c\right)=P\left(2\ln\frac{P\left(\tilde{Y}\mid\hat{\theta}\right)}{P\left(\tilde{Y}\mid\theta\right)}<-2\ln c\right)=P\left(\chi_1^2<-2\ln c\right). \tag{W21}$$

Zatem jeśli dla jakiegoś $0<(1-\alpha)<1$ wybierzemy parametr obcięcia:

$$c=e^{-\frac{1}{2}\chi^2_{1,(1-\alpha)}}, \tag{W22}$$

gdzie $\chi^2_{1,(1-\alpha)}$ jest kwantylem rzędu $100(1-\alpha)\%$ rozkładu χ-kwadrat, to spełnienie przez θ związku:

$$P\left(\frac{P\left(\tilde{Y}\mid\theta\right)}{P\left(\tilde{Y}\mid\hat{\theta}\right)}>c\right)=P\left(\chi_1^2<\chi^2_{1,(1-\alpha)}\right)=1-\alpha \tag{W23}$$

oznacza, że przyjęcie wartości c zgodnej z (W22) daje zbiór możliwych wartości parametru θ:

$$\left\{\theta,\frac{P\left(\tilde{Y}\mid\theta\right)}{P\left(\tilde{Y}\mid\hat{\theta}\right)}>c\right\}, \tag{W24}$$

nazywany $100\,(1-\alpha)\%$-owym (*wiarygodnościowym*) *przedziałem ufności*. Jest on odpowiednikiem wyznaczonego na poziomie ufności $(1-\alpha)$ częstotliowściowego przedziału ufności dla θ. Dla analizowanego przypadku rozkładu normalnego z estymacją skalarnego parametru θ oczekiwanego poziomu zjawiska, otrzymujemy po skorzystaniu z wzoru (W22) wartość parametru obcięcia równego $c=0.15$ lub $c=0.04$ dla odpowiednio 95%-owego $(1-\alpha=0.95)$ bądź 99%-owego $(1-\alpha=0.99)$ przedziału ufności. Tak więc w przypadku, *gdy przedział wiarygodności da się wyskalować rozkładem prawdopodobieństwa, parametr obcięcia c posiada własność wielkości obserwowanej, interpretowanej częstotliwościowo poprzez związek z poziomem ufności*.

Zwróćmy uwagę, że chociaż konstrukcje częstotliwościowego i wiarygodnościowego przedziału ufności są różne, to *ich losowość wynika* w obu przypadkach *z rozkładu prawdopodobieństwa estymatora* $\hat{\theta}$.

Ćwiczenie: W oparciu o powyższe rozważania wyznaczyć, korzystając z (W16) ogólną postać przedziału wiarygodności dla skalarnego parametru θ rozkładu normalnego.

W1.2.2 Rozkłady regularne

Dla zmiennych o innym rozkładzie niż rozkład normalny, statystyka Wilka W ma w ogólności inny rozkład niż χ^2 [6]. Jeśli więc zmienne nie mają dokładnie rozkładu normalnego lub dysponujemy za małą próbką by móc odwoływać się do (wynikających z twierdzeń granicznych) rozkładów granicznych dla estymatorów parametrów, wtedy związek (W20) (więc i (W22)) daje jedynie przybliżone wyskalowanie przedziału wiarygodności rozkładem χ^2.

Jednakże w przypadkach wystarczająco *regularnych rozkładów*, zdefiniowanych jako takie, w których możemy zastosować przybliżenie kwadratowe:

$$\ln\frac{P(y\mid\theta)}{P(y\mid\hat{\theta})} \approx -\frac{1}{2}\mathbf{iF}(\hat{\theta})(\hat{\theta}-\theta)^2 , \qquad \text{(W25)}$$

powyższe rozumowanie oparte o wyskalowanie wiarygodności rozkładem χ_1^2 jest w przybliżeniu słuszne. Wielkość $\mathbf{iF}(\hat{\theta})$, która pojawiła się powyżej jest *obserwowaną* informacją Fishera, a powyższa formuła stanowi poważne narzędzie w analizie doboru modeli [1,6]. Można powiedzieć, że cały skrypt [1] koncentruje się na analizie zastosowania (wartości oczekiwanej) tego wyrażenia i jego uogólnień. Do sprawy tej wrócimy dalej.

Przykład: Rozważmy przypadek parametru skalarnego θ w jednym eksperymencie ($N=1$) ze zmienną Y posiadającą rozkład Bernoulliego z $m=15$. W wyniku pomiaru zaobserwowaliśmy wartość $Y = \boldsymbol{y} = 3$. Prosta analiza pozwala wyznaczyć wiarygodnościowy przedział ufności dla parametru θ. Ponieważ przestrzeń V_θ parametru θ wynosi $V_\theta = (0,1)$, zatem łatwo pokazać, że dla $c = 0{,}01$, $c = 0{,}1$ oraz $c = 0{,}5$ miałby on realizację odpowiednio $(0{,}019;0{,}583)$, $(0{,}046;0{,}465)$ oraz $(0{,}098;0{,}337)$. Widać, że wraz ze wzrostem wartości c, przedział wiarygodności zacieśnia się wokół wartości oszacowania punktowego $\hat{\theta} = y/m = 1/5$ parametru θ i nic dziwnego, bo wzrost wartości c oznacza akceptowanie jako

możliwych do przyjęcia tylko takich *modelowych wartości parametru* θ, które gwarantują wystarczająco wysoką wiarygodność próbki.

Powyższy przykład pozwala nabyć pewnej intuicji co do sensu stosowania ilorazu funkcji wiarygodności. Mianowicie po otrzymaniu w pomiarze określonej wartości y/m oszacowującej parametr θ, jesteśmy skłonni preferować model z taką wartością parametru θ, która daje większą wartość (logarytmu) ilorazu wiarygodności $P(y|\theta)/P(y|\hat{\theta})$. Zgodnie z podejściem statystyki klasycznej *nie oznacza to jednak*, że uważamy, że parametr θ ma jakiś rozkład. Jedynie wobec niewiedzy co do modelowej (populacyjnej) wartość parametru θ preferujemy ten model, który daje większą wartość ilorazu wiarygodności w próbce.

W1.2.3 Weryfikacja hipotez z wykorzystaniem ilorazu wiarygodności

Powyżej wykorzystaliśmy funkcję wiarygodności do *estymacji wartości parametru* Θ. Funkcję wiarygodności można również wykorzystać w drugim typie wnioskowania statystycznego, tzn. w *weryfikacji hipotez statystycznych*.

Rozważmy prostą hipotezę zerową $H_0 : \Theta = \Theta_0$ wobec złożonej hipotezy alternatywnej $H_1 : \Theta \neq \Theta_0$. W celu przeprowadzenia *testu statystycznego* wprowadźmy unormowaną funkcję wiarygodności:

$$\frac{P(y|\Theta_0)}{P(y|\hat{\Theta})}, \tag{W26}$$

skonstruowaną przy założeniu prawdziwości hipotezy zerowej. Hipotezę zerowa H_0 odrzucamy na rzecz hipotezy alternatywnej, jeśli jej wiarygodność $P(y|\Theta_0)$ jest ”za mała”. Sugerowałoby to, że złożona hipoteza alternatywna H_1 zawiera pewną hipotezę prostą, która jest lepiej poparta przez dane otrzymane w próbce, niż hipoteza zerowa.

Jak o tym wspomnieliśmy powyżej, np. 5%-owe obcięcie c w zagadnieniu estymacyjnym, samo w sobie nie mówi nic o frakcji liczby przedziałów wiarygodności pokrywających nieznaną wartość szacowanego parametru. Potrzebne jest wyskalowanie ilorazu wiarygodności. Również dla weryfikacji hipotez skalowanie wiarygodności jest istotne. Stwierdziliśmy, że takie skalowanie jest możliwe wtedy gdy mamy do czynienia z

jednoparametrowym przypadkiem rozkładu Gaussa, a przynajmniej z przypadkiem wystarczająco regularnym.

Empiryczny poziom istotności: W przypadku jednoparametrowego, regularnego problemu z $(\Theta \equiv (\theta)_{n=1}^{N})$ jak w Przykładzie z Rozdziału W1.2.1, skalowanie poprzez wykorzystanie statystki Wilka służy otrzymaniu empirycznego poziomu istotności p. Ze związku (W20) otrzymujemy wtedy przybliżony (a dokładny dla rozkładu normalnego) *empiryczny poziom istotności*:

$$\begin{aligned} p &\approx P\left(\frac{P\left(\tilde{Y} \mid \hat{\theta}\right)}{P\left(\tilde{Y} \mid \theta_0\right)} \geq \frac{P\left(y \mid \hat{\theta}_{obs}\right)}{P\left(y \mid \theta_0\right)} \right) = P\left(2\ln \frac{P\left(\tilde{Y} \mid \hat{\theta}\right)}{P\left(\tilde{Y} \mid \theta_0\right)} \geq -2\ln c_{obs} \right) \\ &= P\left(\chi_1^2 \geq -2\ln c_{obs}\right), \quad \text{g}dzie \quad c_{obs} \equiv \frac{P\left(y \mid \theta_0\right)}{P\left(y \mid \hat{\theta}_{obs}\right)}, \end{aligned} \tag{W27}$$

przy czym $\hat{\theta}_{obs}$ jest wartością estymatora MNW $\hat{\theta}$ wyznaczoną w obserwowanej (obs) próbce y. Powyższe określenie empirycznego poziomu istotności p oznacza, że w przypadku wystarczająco regularnego problemu [6], istnieje typowy związek pomiędzy prawdopodobieństwem (W23), a empirycznym poziomem istotności p, podobny do związku jaki istnieje pomiędzy poziomem ufności $1-\alpha$, a poziomem istotności α w analizie częstotliwościowej. I tak, np. w przypadku jednoparametrowego rozkładu normalnego możemy wykorzystać wartość empirycznego poziomu istotności p do stwierdzenia, że gdy $p \leq \alpha$ to hipotezę H_0 odrzucamy na rzecz hipotezy H_1, a w przypadku $p > \alpha$ nie mamy podstawy do odrzucenia H_0.

Problem błędu pierwszego i drugiego rodzaju: Jednakże podobne skalowanie ilorazu wiarygodności okazuje się być znacznie trudniejsze już chociażby tylko w przypadku dwuparametrowego rozkładu normalnego, gdy obok θ estymujemy σ^2 [6]. Wtedy określenie co oznacza sformułowanie ,,zbyt mała" wartość c jest dość dowolne i zależy od rozważanego problemu lub wcześniejszej wiedzy wynikającej z innych źródeł niż prowadzone statystyczne wnioskowanie. Wybór dużego parametru obcięcia c spowoduje, że istnieje większe prawdopodobieństwo popełnienia *błędu pierwszego rodzaju* polegającego na odrzuceniu hipotezy zerowej w przypadku, gdy jest ona prawdziwa. Wybór małego c spowoduje zwiększenie prawdopodobieństwa popełnienia *błędu drugiego rodzaju*, tzn. przyjęcia hipotezy zerowej w sytuacji, gdy jest ona błędna.

W1.3 MNW w analizie regresji

Analiza zawarta w całym Rozdziale W1.3 oparta jest na przedstawieniu metody MNW w analizie regresji klasycznej podanym w [4,7].

W metodzie regresji klasycznej, estymatory parametrów strukturalnych modelu regresji są otrzymane arytmetyczną metodą najmniejszych kwadratów (MNK). Zmienne objaśniające $X_n = x_n$, $n = 1,..., N$, nie mają wtedy charakteru stochastycznego, co oznacza, że eksperyment jest ze względu na nie kontrolowany.

MNK polega na minimalizacji sumy kwadratów odchyleń obserwowanych wartości zmiennej objaśnianej (tzw. odpowiedzi) od ich wartości teoretycznych spełniających równanie regresji. MNK ma znaczenie probabilistyczne tylko w przypadku analizy standardowej, gdy zmienna objaśniana Y ma rozkład normalny. Jej estymatory pokrywają się wtedy z estymatorami MNW. Pokażemy, że tak się sprawy mają.

Załóżmy, że zmienne $Y_1, Y_2,..., Y_N$ odpowiadające kolejnym wartościom zmiennej objaśniającej, x_1, $x_2,..., x_N$, są względem siebie niezależne i mają rozkład normalny ze średnią $\mu_n = E(Y|x_n) = E(Y_n)$ zależną od wariantu zmiennej objaśniającej x_n, oraz taką samą wariancję $\sigma^2(Y_n) = \sigma^2(Y)$.

Funkcja wiarygodności próbki $(y_1, y_2,..., y_N)$ dla normalnego klasycznego modelu regresji z parametrem $\Theta = \mu \equiv (\mu_n)_{n=1}^N$, ma postać:

$$\begin{aligned} P(\mu) \equiv P(y \mid \mu) &= \prod_{n=1}^{N} f(y_n \mid \mu_n) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_n - \mu_n)^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \mu_n)^2\right\}, \end{aligned} \tag{W28}$$

gdzie $f(y_n \mid \mu_n)$, $n = 1,2..., N$, są punktowymi rozkładami gestości prawdopodobieństwa Gaussa. Widać, że maksymalizacja $P(\mu)$ ze względu na $(\mu_n)_{n=1}^N$ pociąga za sobą minimalizację sumy kwadratów reszt[5] (SKR):

$$SKR = \sum_{n=1}^{N}(y_n - \mu_n)^2 , \tag{W29}$$

[5] SSE w literaturze angielskiej.

gdzie $\mu_n = E(Y|x_n)$ jest *postulowanym modelem regresji*. Zatem w standardowej, klasycznej analizie regresji, estymatory MNW pokrywają się z estymatorami MNK. Widać, że procedura minimalizacji dla SKR prowadzi do liniowej w Y_n postaci estymatorów $\hat{\mu}_n$ parametrów μ_n.

Problem z nieliniowym układem równań wiarygodności: Jednak rozwiązanie układu równań wiarygodności (W6) jest zazwyczaj nietrywialne. Jest tak, gdy otrzymany w wyniku ekstremizacji układ algebraicznych równań wiarygodności dla estymatorów jest nieliniowy, co w konsekwencji oznacza, że możemy nie otrzymać ich w zwartej analitycznej postaci. Przykładem może być analiza regresji Poissona, w której do rozwiązania równań wiarygodności wykorzystujemy metody iteracyjne. W takich sytuacjach wykorzystujemy na ogół jakiś program komputerowy do analizy statystycznej, np. zawarty w pakiecie SAS. Po podaniu postaci funkcji wiarygodności, program komputerowy dokonuje jej maksymalizacji rozwiązując układ (W6) np. metodą Newton-Raphson'a [6,7], wyznaczając numerycznie wartości estymatorów parametrów modelu.

Testy statystyczne: Logarytm ilorazu wiarygodności jest również wykorzystywany w analizie regresji do przeprowadzania testów statystycznych przy weryfikacji hipotez o nie występowaniu braku dopasowania modelu mniej złożonego, tzw. ”niższego”, o mniejszej liczbie parametrów, w stosunku do bardziej złożonego modelu ”wyższego”, posiadającego większą liczbę parametrów. Statystyka wykorzystywana do tego typu testów ma postać [4,6,7]:

$$-2\ln\frac{P\left(\tilde{Y}\mid\hat{\Theta}_1\right)}{P\left(\tilde{Y}\mid\hat{\Theta}_2\right)} \qquad \text{(W30)}$$

gdzie $P\left(\tilde{Y}\mid\hat{\Theta}_1\right)$ jest maksymalizowaną wartością funkcji wiarygodności dla modelu mniej złożonego, a $P\left(\tilde{Y}\mid\hat{\Theta}_2\right)$ dla modelu bardziej złożonego. Przy prawdziwości hipotezy zerowej H_0 o braku konieczności rozszerzania modelu niższego do wyższego, statystyka (W30) ma asymptotycznie rozkład χ^2 z liczbą stopni swobody równą różnicy liczby parametrów modelu wyższego i niższego.

Analogia współczynnika determinacji: Maksymalizowana wartość funkcji wiarygodności zachowuje się podobnie jak *współczynnik determinacji* R^2 [4,7], tzn. rośnie wraz ze wzrostem liczby parametrów w modelu, zatem wielkość pod logarytmem należy do

przedziału $(0,1)$ i statystyka (W30) przyjmuje wartości z przedziału $(0,+\infty)$. Stąd (asymptotycznie) zbiór krytyczny dla H_0 jest prawostronny. Im lepiej więc model wyższy dopasowuje się do danych empirycznych w stosunku do modelu niższego, tym większa jest wartość statystyki ilorazu wiarygodności (W30) i większa szansa, że wpadnie ona w przedział odrzuceń hipotezy zerowej H_0, który leży w prawym ogonie wspomnianego rozkładu χ^2 [4,7].

W1.3.1 Dewiancja jako miara dobroci dopasowania. Rozkład Poissona

Rozważmy zmienną losową Y posiadającą rozkład Poissona. Rozkład ten jest wykorzystywany do modelowania zjawisk związanych z rzadko zachodzącymi zdarzeniami, jak na przykład z liczbą rozpadających się niestabilnych jąder w czasie t. Ma on postać:

$$p(Y = y \mid \mu) = \frac{\mu^y e^{-\mu}}{y!}, \quad oraz \quad y = 0,1,...,\infty, \tag{W31}$$

gdzie μ jest parametrem rozkładu. Zmienna losowa podlegająca rozkładowi Poissona może przyjąć tylko nieujemną wartość całkowitą. Rozkład ten można wyprowadzić z rozkładu dwumianowego, bądź wykorzystując rozkłady Erlanga i wykładniczy [2].
Na przykład, zgodnie z (W31) prawdopodobieństwo, że Y przyjmuje wartość $y = 7$ wynosi:

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

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

Związek wariancji z wartością oczekiwaną rozkład Poissona: Rozkład Poissona posiada pewną interesującą właściwość statystyczną, mianowicie jego wartość oczekiwana, wariancja i trzeci moment centralny są równe parametrowi rozkładu μ:

$$E(Y) = \sigma^2(Y) = \mu_3 = \mu. \tag{W32}$$

Aby pokazać dwie pierwsze równości w (W32) skorzystajmy bezpośrednio z definicji odpowiednich momentów, otrzymując:

$$E(Y) = \sum_{y=0}^{\infty} y \cdot p(Y = y \mid \mu) = \sum_{y=0}^{\infty} y \cdot \frac{\mu^y e^{-\mu}}{y!} = e^{-\mu} \sum_{y=1}^{\infty} \frac{\mu^y}{(y-1)!}$$
$$= e^{-\mu} \mu \sum_{y=1}^{\infty} \frac{\mu^{y-1}}{(y-1)!} = e^{-\mu} \mu \sum_{l=0}^{\infty} \frac{\mu^l}{l!} = e^{-\mu} \mu e^{\mu} = \mu , \qquad \text{(W33)}$$

oraz, korzystając z (W33):

$$\sigma^2(Y) = E(Y^2) - [E(Y)]^2 = E(Y^2) - \mu^2 = \sum_{y=0}^{\infty} y^2 \cdot p(Y = y \mid \mu) - \mu^2$$
$$= \sum_{y=0}^{\infty} y^2 \cdot \frac{\mu^y e^{-\mu}}{y!} - \mu^2 = e^{-\mu} \sum_{y=1}^{\infty} y \frac{\mu^y}{(y-1)!} - \mu^2 = e^{-\mu} \mu \sum_{l=0}^{\infty} (l+1) \frac{\mu^l}{l!} - \mu^2 \qquad \text{(W34)}$$
$$= e^{-\mu} \mu \left[\sum_{l=0}^{\infty} l \frac{\mu^l}{l!} + e^{\mu} \right] - \mu^2 = e^{-\mu} \mu \left[e^{\mu} \mu + e^{\mu} \right] - \mu^2 = (\mu^2 + \mu) - \mu^2 = \mu .$$

Uwaga: *Zatem otrzymaliśmy ważną własność rozkładu Poissona*, która mówi, że stosunek dyspersji σ do wartości oczekiwanej $E(Y)$ maleje pierwiastkowo wraz ze wzrostem poziomu zmiennej Y opisanej tym rozkładem:

$$\frac{\sigma}{E(Y)} = \frac{1}{\sqrt{\mu}} \ . \qquad \text{(W35)}$$

Fakt ten oznacza z założenia *inne zachowanie się odchylenia standardowego* w modelu regresji Poissona niż w klasycznym modelu regresji normalnej (w którym zakładamy jednorodność wariancji zmiennej objaśnianej w różnych wariantach zmiennej objaśniającej).

Ćwiczenie: Pokazać (W32) dla trzeciego momentu.

Przyczyna nielosowej zmiany wartości zmiennej objaśnianej: Rozważmy model regresji dla zmiennej objaśnianej Y posiadającej rozkład Poissona. Zmienne Y_n, $n = 1,2,...,N$ posiadają więc również rozkład Poissona i zakładamy, że są *parami wzajemnie niezależne*. Niech X jest zmienną objaśniającą (tzw. czynnikiem) kontrolowanego eksperymentu, w którym X nie jest zmienną losową, ale *jej zmiana*, jest rozważana jako możliwa przyczyna warunkująca *nielosową zmianę wartości zmiennej* Y.

Gdy czynników $X_1, X_2, ...X_k$ jest więcej, wtedy dla każdego punktu n próby podane są wszystkie ich wartości:

$$x_{1n}, x_{2n}, ...x_{kn} , \quad gdzie \quad n = 1,2,...,N , \qquad \text{(W36)}$$

gdzie pierwszy indeks w x_{in}, $i = 1,2,...,k$, numeruje zmienną objaśniającą.

Brak możliwości eksperymentalnej separacji podstawowego kanału n: Niech $x_n = (x_{1n}, x_{2n},..., x_{kn})$ oznacza zbiór wartości jednego wariantu zmiennych $(X_1, X_2,..., X_k)$, tzn. dla jednej konkretnej podgrupy n. Zwróćmy uwagę, że *indeks próby* n numeruje podgrupę, co oznacza, że w pomiarze wartości Y_n nie ma możliwości eksperymentalnego sięgnięcia "w głąb" indeksu n - tego kanału, tzn. do rozróżnienia wpływów na wartość y_n płynących z różnych "pod-kanałów" i, gdzie $i = 1,2,..., k$.

W1.3.2 Model podstawowy

Zakładając brak zależności zmiennej Y od czynników $X_1, X_2,...X_k$, rozważa się tzw. *model podstawowy*. Dla rozkładu (W31) i próby $\tilde{Y} \equiv (Y_n)_{n=1}^N$, funkcja wiarygodności przy parametrze $\Theta = \mu \equiv (\mu_n)_{n=1}^N$, ma postać:

$$P(\tilde{Y} \mid \mu) = \prod_{n=1}^{N} \frac{\mu_n^{Y_n} e^{-\mu_n}}{Y_n!} = \frac{\left(\prod_{n=1}^{N} \mu_n^{Y_n}\right) \exp\left(-\sum_{n=1}^{N} \mu_n\right)}{\prod_{n=1}^{N} Y_n!}, \qquad \text{(W37)}$$

jest więc wyrażona jako funkcja wektorowego parametru $\mu \equiv (\mu_n)_{n=1}^N$, gdzie każdy z parametrów $\mu_n = E(Y_n)$ jest parametrem skalarnym. N jest równocześnie liczebnością zbioru danych, która może być liczbą podgrup, komórek lub kategorii, oraz liczbą parametrów modelu podstawowego występującą w wiarygodności (W37).

Rozważmy układ równań MNW:

$$\frac{\partial}{\partial \mu_n} \left[\ln P(\tilde{Y} \mid \mu)\right] = 0, \quad n = 1,2,..., N. \qquad \text{(W38)}$$

Dla funkcji wiarygodności (W37) otrzymujemy:

$$\ln P(\tilde{Y} \mid \mu) = \sum_{n=1}^{N} Y_n \ln \mu_n - \sum_{n=1}^{N} \mu_n - \sum_{n=1}^{N} \ln Y_n!. \qquad \text{(W39)}$$

Zatem rozwiązanie układu (W38) daje:

$$\mu_n = \hat{\mu}_n = Y_n, \quad n = 1,2,..., N, \qquad \text{(W40)}$$

jako estymatory modelu postawowego. Zatem funkcja wiarygodności (W37) modelu podstawowego przyjmuje w punkcie μ zadanym przez estymatory (W40) wartość maksymalną:

$$P\left(\tilde{Y} \mid \hat{\mu}\right) = \frac{\left(\prod_{n=1}^{N} Y_n^{Y_n}\right)\exp\left(-\sum_{n=1}^{N} Y_n\right)}{\prod_{n=1}^{N} Y_n!}, \qquad \text{(W41)}$$

gdzie zastosowano oznaczenie $\hat{\mu} = \left(\hat{\mu}_1, \hat{\mu}_2, ...\hat{\mu}_N\right)$.

W1.3.3. Analiza regresji Poissona.

Niech zmienna zależna Y reprezentuje liczbę zliczeń badanego zjawiska (np. przypadków awarii określonego zakupionego sprzętu), otrzymaną dla każdej z N podgrup (np. klienckich). Każda z tych podgrup wyznaczona jest przez komplet wartości zmiennych objaśniających $X \equiv \left(X_1, X_2, ..., X_k\right) = x \equiv \left(x_1, x_2, ..., x_k\right)$ (np. wiek, poziom wykształcenia, cel nabycia sprzętu). Zmienna Y_n określa liczbę zliczeń zjawiska w n-tej podgrupie, $n = 1,2,..., N$. W konkretnej próbce $(Y_n)_{n=1}^N = (y_n)_{n=1}^N$.

Określenie modelu regresji Poissona: Rozważmy następujący model regresji Poissona:

$$\mu_n \equiv E\left(Y_n\right) = \ell_n r\left(x_n, \beta\right), \quad n = 1,2,..., N, \qquad \text{(W42)}$$

opisujący zmianę wartości oczekiwanej liczby zdarzeń Y_n (dla rozkładu Poissona) wraz ze zmianą *wariantu* $x_n = \left(x_{1n}, x_{2n}, ..., x_{kn}\right)$.

Funkcja regresji po prawej stronie (W42) ma dwa czynniki. Czynnik funkcyjny funkcji regresji, $r\left(x_n, \beta\right)$, opisuje *tempo zdarzeń* określanych mianem porażek (np. awarii) w n-tej podgrupie (tzn. jest *częstością* tego zjawiska), skąd $r\left(x_n, \beta\right) > 0$, gdzie $\beta \equiv \left(\beta_0, \beta_1, ..., \beta_k\right)$ jest zbiorem nieznanych parametrów tego modelu regresji. Natomiast czynnik ℓ_n jest współczynnikiem określającym *dla każdej n-tej podgrupy* (np. klientów) *skumulowany czas prowadzenia badań kontrolnych dla wszystkich jednostek tej podgrupy*.

Ponieważ funkcja regresji[6] $r\left(x_n, \beta\right)$ przedstawia typową liczbę porażek na jednostkę czasu, zatem nazywamy ją *ryzykiem*.

[6] Czynnik $r\left(x_n, \beta\right)$ nazywany dalej funkcją regresji, chociaż właściwie nazwa ta odnosi się do całej $E\left(Y_n\right)$.

Uwaga o postaci funkcji regresji: Funkcję $r(x_n,\beta)$ można zamodelować na różne sposoby [6]. Wprowadźmy oznaczenie:

$$\lambda_n^* \equiv \beta_0 + \sum_{j=1}^{k} \beta_j x_{jn} \,. \tag{W43}$$

Funkcja regresji $r(x_n,\beta)$ ma różną postać w zależności od typu danych. Może mieć ona postać charakterystyczną dla regresji liniowej (wielokrotnej), $r(x_n,\beta) = \lambda_n^*$, którą stosujemy szczególnie wtedy gdy zmienna Y ma *rozkład normalny*. Postać $r(x_n\beta) = 1/\lambda_n^*$ jest stosowana w analizie z danymi pochodzącymi z *rozkładu eksponentialnego*, natomiast $r(x_n,\beta) = 1/(1+\exp(-\lambda_n^*))$ w modelowaniu regresji logistycznej dla opisu zmiennej *dychotomicznej* [4,6].

Postać funkcji regresji użyteczna w regresji Poissona jest następująca:

$$r(x_n,\beta) = \exp(\lambda_n^*)\,, \quad \lambda_n^* = \beta_0 + \sum_{j=1}^{k} \beta_j x_{jn} \,. \tag{W44}$$

Ogólniej mówiąc analiza regresji odnosi się do modelowania wartości oczekiwanej zmiennej zależnej (objaśnianej) jako funkcji pewnych czynników. Postać funkcji wiarygodności stosowanej do estymacji współczynników regresji β odpowiada założeniom dotyczącym rozkładu zmiennej zależnej. Tzn. zastosowanie konkretnej funkcji regresji $r(x_n,\beta)$, np. jak w (W44), wymaga określenia postaci funkcji częstości $r(x_n,\beta)$, zgodnie z jej postacią dobraną do charakteru losowej zmiennej Y przy której generowane są dane w badanym zjawisku. Na ogół przy konstrukcji $r(x_n,\beta)$ pomocna jest uprzednia wiedza dotyczącą relacji między rozważanymi zmiennymi.

Funkcja wiarygodności dla analizy regresji Poissona: Ponieważ Y_n ma rozkład Poissona (W31) ze średnią μ_n, $p(Y_n \mid \mu_n) = \dfrac{\mu_n^{Y_n}}{Y_n!} e^{-\mu_n}$, $n = 1,2,...,N$, zatem dane $Y_n = 0,1,...,\infty$ dla określonego $n = 1,2,...,N$ są generowane z rozkładów warunkowych:

$$p(Y_n \mid \beta) = \frac{[\ell_n r(x_n,\beta)]^{Y_n}}{Y_n!} e^{-\ell_n r(x_n,\beta)} \,, \tag{W45}$$

wokół funkcji regresji, (W42), $\mu_n = \ell_n r(x_n,\beta)$, dla $n = 1,2,...,N$. Funkcja wiarygodności dla analizy regresji Poissona ma więc postać:

$$P(\tilde{Y} \mid \beta) = \prod_{n=1}^{N} p(Y_n \mid \beta) = \prod_{n=1}^{N} \frac{(\ell_n r(x_n, \beta))^{Y_n} e^{-\ell_n r(x_n, \beta)}}{Y_n!}$$
$$= \frac{\prod_{n=1}^{N} (\ell_n r(x_n, \beta))^{Y_n} \exp\left[-\sum_{n=1}^{N} \ell_n r(x_n, \beta)\right]}{\prod_{n=1}^{N} Y_n!}. \qquad \text{(W46)}$$

Aby w praktyce posłużyć się funkcją regresji $r(x_n, \beta)$ będącą określoną funkcją zmiennej $\lambda_n^* = \beta_0 + \sum_{j=1}^{k} \beta_j x_{jn}$, parametry $\beta_0, \beta_1, ..., \beta_k$ muszą być oszacowane. Estymatory MNW, $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$, tych parametrów otrzymuje się rozwiązując $k+1$ równań wiarygodności:

$$\frac{\partial}{\partial \beta_j} \ln P(\tilde{Y} \mid \beta) = 0, \quad j = 0,1,2,...,k. \qquad \text{(W47)}$$

W przypadku regresji Poissona $P(\tilde{Y} \mid \beta)$ jest określona zgodnie z (W46).

Algorytmy IRLS: Zauważmy, że dla rozkładu Poissona zachodzi zgodnie z (W32) oraz (W42), $\sigma^2(Y_n) = E(Y_n) = \ell_n r(x_n, \beta)$, *co oznacza, że wariancja* $\sigma^2(Y_n)$ *zmiennej objaśnianej nie jest stała lecz zmienia się jako funkcja* ℓ_n *oraz* x_n*, wchodząc w analizę z różnymi wagami wraz ze zmianą* n. Na fakt ten zwróciliśmy już uwagę przy okazji związku (W35). Ponieważ układ równań wiarygodności (W47) jest na ogół rozwiązywany iteracyjnymi metodami numerycznymi [4], a wariancja $\sigma^2(Y_n)$ jest również funkcją β, zatem na każdym kroku procesu iteracyjnego *wagi* te zmieniają się jako funkcja zmieniających się składowych estymatora $\hat{\beta}$. Algorytmy takiej analizy określa się ogólnym mianem *algorytmów najmniejszych kwadratów*[7] *iteracyjnie ważonych* (IRLS[8]) [6, 4]. Nazwa ta pozostała jedynie z powodu „pierwszeństwa" MNK, ale ogólnie nie odnosi się do MNK, która ma probabilistyczne znaczenie tylko gdy zmienne Y_j mają rozkład normalny.

Uwaga o programach: Różne programy do analiz statystycznych, w tym SAS wykorzystujący procedurę PROC GENMOD, mogą być użyte do znajdowania estymatorów $\hat{\beta}$ MNW dla funkcji wiarygodności (W46). Również *obserwowana macierz kowariancji*

[7] Należy jednak pamiętać, że zwrotu "najmniejszych kwadratów" nie należy tu brać dosłownie, gdyż metoda najmniejszych kwadratów ma sens jedynie wtedy, gdy rozkład zmiennej Y jest normalny (por. Rozdział W1.3).

[8] *iteratively reweighted least squares*

estymatorów[9] oraz miary dobroci dopasowania modelu, takie jak omówiona dalej dewiancja, mogą być otrzymane przy użyciu powyżej wspomnianych programów.

W1.3.3.1 Test statystyczny dla doboru modelu w regresji Poissona

Uwaga o większej wiarygodności modelu podstawowego: Maksymalna wartość funkcji wiarygodności $P(y|\mu)$ wyznaczona w oparciu o (W41) będzie, dla każdego zbioru danych i dla liczby parametrów $k+1<N$, większa niż otrzymana przez maksymalizację funkcji wiarygodności (W46). Jest tak, ponieważ w wyrażeniu (W41) na funkcję wiarygodności modelu podstawowego *nie narzuca się żadnych ograniczeń na postać* μ_n, natomiast (W46) wymaga aby $\mu_n = \ell_n r(x_n,\beta)$.

Pomyśl o tym tak: Model podstawowy dopasowuje się do danych, w każdym punkcie z osobna, leżąc zgodnie z (W40) maksymalnie blisko tych danych, natomiast MNW dla modelu regresji $\mu_n \equiv E(Y_n) = \ell_n r(x_n,\beta)$, $n=1,2,...,N$, (W42), wyznacza krzywą regresji przechodzącą pomiędzy punkami pomiarowymi.

Hipoteza zerowa o nie występowaniu braku dopasowania w modelu niższym: Zgodnie z powyższym zdaniem, analizę doboru modelu regresji można rozpocząć od postawienia hipotezy zerowej wobec alternatywnej. W hipotezie zerowej wyróżnimy proponowany model regresji. Wybór modelu badanego oznacza wybór funkcji wiarygodności (W46) z nim związanej.

Stawiamy więc hipotezę zerową:

$$\mathrm{H}_0 : \mu_n = \ell_n r(x_n,\beta), \quad n=1,2,...,N, \tag{W49}$$

która odpowiada wyborowi modelu z funkcją wiarygodności (W46), wobec hipotezy alternatywnej:

$$\mathrm{H}_A : \ \mu_n \text{ n}ie\ ma\ ograniczonej\ \ postaci, n=1,2,...,N, \tag{W50}$$

która odpowiada wyborowi modelu podstawowego zawierającego tyle parametrów μ_n ile jest punktów pomiarowych, tzn. *N*, z funkcją wiarygodności (W41).

---

[9] Obserwowana macierz (wariancji-) kowariancji $\hat{V}(\hat{\beta})$ estymatorów $\hat{\beta}$ MNW jest zdefiniowana jako odwrotność macierzy obserwowanej informacji Fishera [6,1] (por. (D15)):

$$\hat{V}(\hat{\beta}) := \mathbf{iF}^{-1}(\hat{\beta}). \tag{W48}$$

Niech więc $P(\tilde{Y} \mid \hat{\beta})$ jest maksymalną wartością funkcji wiarygodności określoną jak w (W46). Oznacza to, że w miejsce parametrów $\beta = (\beta_0, \beta_1, ..., \beta_k)$ podstawiono ich estymatory $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)$ wyznaczone przez MNW, jako te które maksymalizują funkcję wiarygodności (W46). Podobnie rozumiemy funkcję wiarygodności $P(\tilde{Y} \mid \hat{\mu})$ modelu podstawowego.

Ponieważ celem każdej analizy jest otrzymanie możliwie najprostszego opisu danych, model $\mu_n = \ell_n r(x_n, \beta)$ zawierający $k+1$ parametrów β, będzie uznany za dobry, jeśli maksymalna wartość funkcji wiarygodności wyznaczona dla niego, będzie prawie tak duża, jak funkcji wiarygodności dla nie niosącego żadnej informacji modelu podstawowego z liczbą parametrów μ_n równą licznie punktów pomiarowych N. Sformułowanie ”prawie tak duża” oznacza, że wartość funkcji wiarygodności $P(y \mid \hat{\beta})$ nie może być istotnie statystycznie mniejsza od $P(y \mid \hat{\mu})$. Zasadniczo powinno to oznaczać, że musimy podać miary pozwalające na określenie statystycznej istotności przy posługiwaniu się intuicyjnym parametrem obcięcia c (Rozdział W1.2). Okazuje się, że dla dużej próby, miary typu (W51), podane poniżej, uzyskują cechy pozwalające na budownie wiarygodnościowych obszarów krytycznych nabywających charakteru standardowego (częstotliwościowego).

Określenie dewiancji: Wprowadźmy *statystykę typu ilorazu wiarygodności*:

$$D(\hat{\beta}) = -2 \ln \left[\frac{P(\tilde{Y} \mid \hat{\beta})}{P(\tilde{Y} \mid \hat{\mu})} \right] \qquad \text{(W51)}$$

nazywaną *dewiancją* (deviance) dla modelu regresji, w tym przypadku dla modelu Poissona z określoną postacią $\mu_n = \ell_n r(x_n, \beta)$. Służy ona do badania dobroci dopasowania modelu z zadaną postacią $\mu_n = \ell_n r(x_n, \beta)$ w stosunku do modelu podstawowego, bez narzuconej postaci na μ_n, tzn. do stwierdzenia, czy $P(y \mid \hat{\beta})$ jest istotnie *mniejsza* od $P(y \mid \hat{\mu})$, co sugerowałoby istotny statystycznie brak dopasowania badanego modelu $\mu_n = \ell_n r(x_n, \beta)$, do danych empirycznych. Jak pokażemy poniżej dewiancja może być rozumiana jako *miara zmienności reszt* (tzn. odchylenia wartości obserwowanych w próbie od wartości szacowanych przez model) *wokół linii regresji,* na której leżą wartości przewidywane $\hat{y}_j$ przez model [1].

Przy prawdziwości hipotezy $H_0 : \mu_n = \ell_n r(x_n, \beta)$, rozkład dewiancji $D(\hat{\beta})$ dla regresji Poissona, można asymptotycznie przybliżyć rozkładem chi-kwadrat (por. dyskusja w [4,6]) z $N-k-1$ stopniami swobody.

Wyznaczenie liczby stopni swobody dewiancji: Podana liczba stopni swobody dewiancji $D(\hat{\beta})$ wynika z następującego rozumowania. Zapiszmy (W51) w postaci:

$$D(\hat{\beta}) + 2\ln P(\tilde{Y} \mid \hat{\beta}) = 2\ln P(\tilde{Y} \mid \hat{\mu}), \tag{W52}$$

co po skorzystaniu z (W46) dla $\beta = \hat{\beta}$ ma postać:

$$D(\hat{\beta}) + 2\sum_{n=1}^{N} \ell_n r(x_n, \hat{\beta}) = 2\ln P(\tilde{Y} \mid \hat{\mu}) + 2\ln\left(\prod_{n=1}^{N} Y_n!\right) - 2\ln\left(\prod_{n=1}^{N} \left(\ell_n r(x_n, \hat{\beta})\right)^{Y_n}\right). \tag{W53}$$

Można zauważyć, że prawa strona tego równania ma N -stopni swobody. Istotnie, ze względu na (W40)[10], $\hat{\mu} \equiv (\hat{\mu}_n) = (Y_n)$, $n = 1,2,...,N$, liczba niezależnych zmiennych po prawej strony powyższego równania, których wartości trzeba określić z eksperymentu, wynosi N. Natomiast drugi składnik po lewej stronie ma liczbę stopni swobody równą $k+1$, co jest liczbą estymatorów parametrów strukturalnych $\hat{\beta}$ modelu regresji, których wartości trzeba określić z eksperymentu. Ponieważ liczba stopni swobody po prawej i lewej stronie równania musi być taka sama, zatem liczba stopni swobody dewiancji $D(\hat{\beta})$ wynosi $N-k-1$.

Decyzja testu statystycznego: Z powyższego wynika, że dla *bardzo dużej* próbki dewiancja $D(\hat{\beta})$ posiada, przy prawdziwości hipotezy $H_0 : \mu_n = \ell_n r(x_n, \beta)$ (W49) w przybliżeniu rozkład chi-kwadrat z $N-k-1$ stopniami swobody. Zatem, przybliżony statystyczny test dobroci dopasowania (tzn. niewystępowania braku dopasowania) modelu $\mu_n = \ell_n r(x_n, \beta)$ do danych w stosunku do modelu podstawowego, może zostać wykonany przez sprawdzenie czy w zaobserwowanej (*obs*) próbce $\tilde{Y} = y$, wartości estymatorów MNW $\hat{\beta} \equiv \hat{\beta}_{obs}$ modelu regresji (W42) oraz $\hat{\mu} \equiv \hat{\mu}_{obs} = y$ modelu podstawowego (W40), dają wartość dewiancji $D(\hat{\beta}) = D(\hat{\beta}_{obs})$:

$$D(\hat{\beta}_{obs}) = -2\ln\left[\frac{P(\tilde{Y} \mid \hat{\beta}_{obs})}{P(\tilde{Y} \mid \hat{\mu}_{obs})}\right] , \tag{W54}$$

[10] W przyjętym przedstawieniu danych jak dla diagramu punktowego, N jest ogólnie liczbą punktów pomiarowych (równą liczbie wariantów czy komórek). Tylko dla modelu podstawowego jest N również liczbą parametrów.

która jest nie mniejsza niż wartość krytyczna w prawym ogonie rozkładu chi-kwadrat z $N-k-1$ stopniami swobody [4]. Przyjęcie przez $D\left(\hat{\beta}_{obs}\right)$ wartości równej lub większej od krytycznej skutkuje odrzuceniem hipotezy zerowej. Alternatywnie, mając wartości $D\left(\hat{\beta}_{obs}\right)$, można policzyć empiryczny poziom istotności $p = \Pr(\chi^2_{N-k-1} \geq D(\hat{\beta}_{obs}))$ i porównać jego wartość z przyjętą (w dziedzinie badań) wartością poziomu istotności α [12]. Gdy $p \leq \alpha$ wtedy odrzucamy hipotezę zerową H_0, która mówi 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. Gdy $p > \alpha$ wtedy nie mamy podstaw do odrzucenia hipotezy zerowej H_0.

Uwaga dotycząca zapisu indeksu *obs*: W dalszej części będziemy pomijać indeks '*obs*' w indeksie wartości estymatora w próbce, za wyjątkiem sytuacji, gdy rozróżnienie pomiędzy estymatorem jako statystyką, a jego realizacją w próbce, nie wynika jasno z kontekstu.

W1.3.3.2 Testy ilorazu wiarygodności

Dewiancje dla hierarchicznych klas modeli mogą służyć do budowy testów stosunku wiarygodności. Zwróćmy szczególnie uwagę na funkcję wiarygodności (W46) zawierającą zbiór parametrów $\beta = (\beta_0, \beta_1, ..., \beta_k)$ z *dewiancją* $D\left(\hat{\beta}\right)$ daną wyrażeniem (W51). Przypuśćmy, że chcemy zweryfikować hipotezę o tym, że $k-r$ (gdzie $0 < r < k$) ostatnich parametrów będących składowymi wektora β jest równych *zeru*.

Hipoteza zerowa, o nieistotności rozszerzenia modelu niższego do wyższego, ma wtedy postać: $$H_0 : \beta_{r+1} = \beta_{r+2} = ... = \beta_k = 0, \qquad \text{(W55)}$$

Hipoteza alternatywna H_A mówi, że przynajmniej jeden z parametrów strukturalnych $\beta_{r+1}, \beta_{r+2}, ..., \beta_k$ jest różny od *zera*.

Funkcja wiarygodności przy prawdziwości hipotezy zerowej H_0, (W53), ma postać taką jak w (W46), tyle, że zastąpiono w niej parametr β parametrem $\beta_{(r)}$:

$$\beta_{(r)} \equiv (\beta_0, \beta_1, ..., \beta_r; 0, 0, ..., 0) \quad \text{gdzie liczba zer wynosi } k-r. \qquad \text{(W56)}$$

Oznaczmy funkcje wiarygodności tego modelu jako $P(\tilde{Y} \mid \beta_{(r)})$, a $\hat{\beta}_{(r)}$ niech będzie estymatorem MNW wektorowego parametru $\beta_{(r)}$, wyznaczonym przez rozwiązanie odpowiadającego mu układu równań wiarygodności (oczywiście dla niezerowych parametrów $\beta_0, \beta_1, ..., \beta_r$). Estymator $\hat{\beta}_{(r)} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_r; 0,0,...,0)$ maksymalizuje funkcję wiarygodności $P\left(\tilde{Y} \mid \beta_{(r)}\right)$.

Test ilorazu wiarygodności dla weryfikacji hipotezy H_0 przeprowadzamy posługując się *statystyką ilorazu wiarygodności*:

$$-2\ln\left[\frac{P\left(\tilde{Y} \mid \hat{\beta}_{(r)}\right)}{P\left(\tilde{Y} \mid \hat{\beta}\right)}\right], \tag{W57}$$

która przy prawdziwości hipotezy zerowej ma asymptotycznie rozkład chi-kwadrat z $k-r$ stopniami swobody, co widać, gdy zapiszemy (W57) jako różnicę dewiancji:

$$-2\ln\left[\frac{P\left(\tilde{Y} \mid \hat{\beta}_{(r)}\right)}{P\left(\tilde{Y} \mid \hat{\beta}\right)}\right] = -2\ln\left[\frac{P\left(\tilde{Y} \mid \hat{\beta}_{(r)}\right)}{P\left(\tilde{Y} \mid \hat{\mu}\right)}\right] + 2\ln\left[\frac{P\left(\tilde{Y} \mid \hat{\beta}\right)}{P\left(\tilde{Y} \mid \hat{\mu}\right)}\right] = D\left(\hat{\beta}_{(r)}\right) - D\left(\hat{\beta}\right), \tag{W58}$$

oraz skorzystamy z podobnej analizy jak dla (W53).

Zatem, przy prawdziwości hipotezy zerowej (W55), którą można zapisać jako $\mathrm{H}_0 : \beta_{r+1} = \beta_{r+2} = ... = \beta_k = 0$, różnica $D(\hat{\beta}_{(r)}) - D(\hat{\beta})$ ma dla dużej próby w przybliżeniu rozkład chi-kwadrat z $k-r$ stopniami swobody. Natomiast **decyzja testu statystycznego** ma w przypadku posługiwania się statystyką testową (W57) analogiczny przebieg jak dla omówionej poprzednio przypadku dewiancji.

Wniosek: Jeśli używamy regresji Poissona do analizowania danych empirycznych, modele tworzące hierarchiczne klasy mogą być porównywane miedzy sobą poprzez wyznaczenie statystyki ilorazu wiarygodności (W57), lub co na jedno wychodzi, poprzez wyznaczenie różnicy (W58) między parami dewiancji dla tych modeli. Należy przy tym pamiętać o wniosku jaki już znamy z analizy dewiancji, że *im model gorzej dopasowuje się do danych empirycznych tym jego dewiancja jest większa.*

W1.3.3.3 Podobieństwo dewiancji do SKR analizy częstotliwościowej

Warunkowe wartości oczekiwane $\mu_n \equiv E(Y_n) = \ell_n r(x_n, \beta)$, $n = 1,2,..., N$, (W42), są w analizie regresji przyjmowane jako teoretyczne przewidywania modelu regresji dla wartości zmiennej objaśnianej Y_n, zwanej odpowiedzią (układu).

W próbie oszacowania $\mu_n \equiv E(Y_n) = \ell_n r(x_n, \beta)$ oznaczamy jako $\hat{Y}_n$. W n-tej komórce jest ono następujące:

$$\hat{Y}_n = \ell_n r(x_n, \hat{\beta}), \quad n = 1,2,..., N, \tag{W59}$$

zgodnie z wyestymowaną postacią modelu regresji. Wykorzystując (W59) w (W46) możemy zapisać dewiancję modelu (W51) następująco:

$$\begin{aligned} D(\hat{\beta}) &= -2\ln\left[\frac{P(\tilde{Y} \mid \hat{\beta})}{P(\tilde{Y} \mid \hat{\mu})}\right] = -2\ln\left[\frac{\prod_{n=1}^{N} \hat{Y}_n^{Y_n} \exp\left(-\sum_{n=1}^{N} \hat{Y}_n\right)}{\prod_{n=1}^{N} Y_n^{Y_n} \exp\left(-\sum_{n=1}^{N} Y_n\right)}\right] \\ &= -2\left[\sum_{n=1}^{N} Y_n \ln \hat{Y}_n - \sum_{n=1}^{N} \hat{Y}_n - \sum_{n=1}^{N} Y_n \ln Y_n + \sum_{n=1}^{N} Y_n\right] \end{aligned} \tag{W60}$$

tzn:

$$D(\hat{\beta}) = 2\sum_{n=1}^{N}\left[Y_n \ln\left(\frac{Y_n}{\hat{Y}_n}\right) - \left(Y_n - \hat{Y}_n\right)\right]. \tag{W61}$$

Podobieństwo D **do** SKR: Powyższa postać dewiancji oznacza, że $D(\hat{\beta})$ zachowuje się w poniższym sensie jak suma kwadratów reszt $SKR = \sum_{n=1}^{N}(Y_n - \hat{Y}_n)^2$ w standardowej wielorakiej regresji liniowej. Otóż, gdy dopasowywany model dokładnie przewiduje obserwowane wartości, tzn. $\hat{Y}_n = Y_n, n = 1,2,..,N$ wtedy, jak SKR w analizie standardowej [4,8], tak $D(\hat{\beta})$ w analizie wiarygodnościowej jest równe zeru [4]. Z drugiej strony wartość $D(\hat{\beta})$ jest tym większa im większa jest różnica między wartościami obserwowanymi Y_n i wartościami przewidywanymi $\hat{Y}_n$ przez oszacowany model.

Asymptotyczna postać D: W analizowanym modelu Y_n, $n = 1,2,..., N$ są niezależnymi zmiennymi Poissona (np. zmiennymi częstości), natomiast wartości $\hat{Y}_n$ są ich przewidywaniami. Nietrudno przekonać się, że gdy wartości przewidywane mają rozsądną

wartość[11], np. $\hat{Y}_n > 3$ oraz $(Y_n - \hat{Y}_n) << Y_n$, $n = 1,2,..,N$ tak, że $(Y_n - \hat{Y}_n)/Y_n << 1$, wtedy wyrażenie w nawiasie kwadratowym w (W61) można przybliżyć przez $(Y_n - \hat{Y}_n)^2/(2Y_n)$, a statystykę (W61) można przybliżyć statystyką o postaci:

$$\chi^2 = \sum_{n=1}^{N} \frac{\left(Y_n - \hat{Y}_n\right)^2}{\hat{Y}_n}, \tag{W62}$$

która (dla dużej próby) ma rozkład chi-kwadrat z $N - k - 1$ stopniami swobody [4].

W1.4 Zasada niezmienniczości ilorazu funkcji wiarygodności

Z powyższych rozważań wynika, że funkcja wiarygodności reprezentuje niepewność dla ustalonego parametru. Nie jest ona jednak gęstością rozkładu prawdopodobieństwa dla tego parametru. Pojęcie takie byłoby całkowicie obce statystyce klasycznej (nie włączając procesów stochastycznych). Inaczej ma się sprawa w tzw. statystyce Bayesowskiej. Aby zrozumieć różnicę pomiędzy podejściem klasycznym i Bayesowskim [12] rozważmy transformację parametru.

Przykład transformacji parametru: Rozważmy eksperyment, w którym dokonujemy jednokrotnego pomiaru zmiennej o rozkładzie dwumianowym (W8). Funkcja wiarygodności ma więc postać $P(\theta) = \binom{m}{x}\theta^x(1-\theta)^{m-x}$. Niech parametr $m = 12$ a w pomiarze otrzymano $x = 9$. Testujemy model, dla którego $\theta = \theta_1 = 3/4$ wobec modelu z $\theta = \theta_2 = 3/10$. Stosunek wiarygodności wynosi:

$$\frac{P(\theta_1 = 3/4)}{P(\theta_2 = 3/10)} = \frac{\binom{m}{x}\theta_1^9(1-\theta_1)^3}{\binom{m}{x}\theta_2^9(1-\theta_2)^3} = 173.774 \tag{W63}$$

Dokonajmy hiperbolicznego wzajemnie jednoznacznego przekształcenia parametru:

$$\psi = 1/\theta. \tag{W64}$$

[11] Zauważmy, że statystyka (W61) może mieć myląco dużą wartość gdy wielkości $\hat{Y}_n$ są bardzo małe.

Funkcja wiarygodności po transformacji parametru ma postać $\tilde{P}(\psi)=\binom{m}{x}(1/\psi)^x(1-1/\psi)^{m-x}$.

Wartości parametru ψ odpowiadające wartościom θ_1 i θ_2 wynoszą odpowiednio $\psi_1 = 4/3$ oraz $\psi_2 = 10/3$. Łatwo sprawdzić, że transformacja (W64) nie zmienia stosunku wiarygodności, tzn.:

$$\frac{\tilde{P}(\psi_1 = 4/3)}{\tilde{P}(\psi_2 = 10/3)} = \frac{P(\theta_1 = 3/4)}{P(\theta_1 = 3/10)} = 173.774. \qquad \text{(W65)}$$

Niezmienniczość stosunku wiarygodności: Zatem widać, że stosunek wiarygodności jest niezmienniczy ze względu na wzajemnie jednoznaczną transformację parametru. Gdyby transformacja parametru była np. transformacją ”logit” $\psi = \ln(\theta/(1-\theta))$ lub paraboliczną $\psi = \theta^2$, to sytuacja także nie uległaby zmianie. Również w ogólnym przypadku transformacji parametru własność *niezmienniczości stosunku wiarygodności* pozostaje słuszna. Oznacza to, że informacja zawarta w próbce jest niezmiennicza ze względu na wybór parametryzacji, tzn. powinniśmy być w takiej samej sytuacji niewiedzy niezależnie od tego jak zamodelujemy zjawisko, o ile różnica w modelowaniu sprowadza się jedynie do transformacji parametru. W omawianym przykładzie powinniśmy równie dobrze móc stosować parametr θ, jak $1/\theta$, θ^2, czy $\ln(\theta/(1-\theta))$.

Uwaga o transformacji parametru w statystyce Bayesowskiej: Natomiast sytuacja ma się zupełnie inaczej w przypadku Bayesowskiego podejścia do funkcji wiarygodności [12], w którym funkcja wiarygodności uwzględnia (Bayesowski) rozkład prawdopodobieństwa $f(\theta \mid x)$ parametru θ. Oznacza to, że Jakobian transformacji $\theta \to \psi$ parametru, modyfikując rozkład parametru, zmienia również funkcję wiarygodności. Zmiana ta zależy od wartości parametru, różnie zmieniając licznik i mianownik w (W63), co niszczy *intuicyjną* własność niezmienniczości ilorazu wiarygodności ze względu na transformację parametru [6].

Do zagadnienia użyteczności własności niezmienniczości ilorazu funkcji wiarygodności przy konstrukcji przedziału wiarygodności, powrócimy w dalszej części.

Dodatek. MNW na przykładzie analizy modeli regresji Poissona

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

Przyjmijmy więc, że zmienna objaśniana jest zmienną losową *Y* przyjmującą wartości *y* zgodnie z rozkładem Poissona (W31), $p(y \mid \mu) = \mu^y e^{-\mu} / y!$, $y = 0,1,...,\infty$. Jak wspomnieliśmy we Wprowadzeniu, rozkład Poissona (W31) jest często używany do modelowania pojawiania się rzadkich zdarzeń, takich jak np. nowych przypadków awarii w pewnej populacji w pewnym okresie czasu albo zajścia określonej liczby wypadków samochodowych w pewnym określonym miejscu w ciągu roku.

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 (nazywanych czynnikiami). Po pierwsze, wyjaśnimy jak konstruować postać modelu regresji Poissona dla tzw. ryzyka względnego i zastosujemy przedstawioną we Wprowadzeniu estymację MNW parametrów modelu. Zastosowanie wnioskowania związanego z weryfikacją hipotez o braku dopasowania w modelu niższym przedstawimy w drugiej kolejności.

D1.1 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 *zaburzenie* 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 wiekiem) 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 parametrem liczby losowań *m*. Związek określający oczekiwaną liczbę sukcesów w *m* losowaniach Bernoulliego $\mu = m\,p$, można zapisać następująco:

$$\mu = (m \cdot \Delta t)\frac{p}{\Delta t} \quad , \tag{D1}$$

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

$$r \equiv \frac{p}{\Delta t} \tag{D2}$$

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 (W42) $\mu_{ij} = \ell_{ij} r_{ij}$ [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:

$$r_{ij} = \frac{\mu_{ij}}{\ell_{ij}} \quad . \tag{D3}$$

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

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

$$\hat{r}_{ij} = \frac{Y_{ij}}{\ell_{ij}} \qquad , \tag{D4}$$

gdzie Y_{ij} jest ilością zaobserwowanych zliczeń defektów silnika dla podgrupy (*i*, *j*), a ℓ_{ij} oznacza zakumulowaną (tzn. sumaryczną) długość czasu wolnego od defektu 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 (D4) zmienna Y_{ij} nie jest w ogólności estymatorem MNW parametru μ_{ij} dla modelu regresji Poissona, chociaż jest tak dla modelu podstawowego.

D1.3.2 Ryzyko względne

Stosunek:

$$R_{Wi} = \frac{r_{i1}}{r_{i0}} \tag{D5}$$

jest parametrem nazywamym *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 (wskaźnika) 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 j-tą (generalną) populację, w której (nielosowe) czynniki X_i, i=1,2,...,p, przyjmują ustalone, im właściwe wartości. W ten sposób liczba wszystkich (pod)populacji wskazanych indeksem j 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.

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* [12]. Ściśle rzecz biorąc, rozkład dwumianowy $B(m, p)$ przechodzi w rozkład Poissona $Poisson(\mu = m\,p)$ zmiennej Y tylko granicznie wtedy, gdy przy liczbie pomiarów *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 [12]. W granicy tej oczekiwana dwumianowa liczba zliczeń „sukcesów" (wartość oczekiwana μ) jest względnie mała w porównaniu z liczbą wszystkich wyników, a rozkład Poissona daje dobre przybliżenie rozkładu dwumianowego dla rzadkich przypadków awarii (które są tu „sukcesami"). 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 satysfakcjonująco spełniają to założenie, gdyż w każdej kategorii wiekowej występuje stosunkowo mały udział względny przypadków awarii w porównaniu do rozmiaru, tzn. liczby wszystkich wyników w odpowiedniej próbce pobranej z 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 Rozdziale D1.7.

D1.6 Równanie regresji Poissona ze zmiennymi ukrytymi

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_k . 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.

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, co oczywiście wpływa na postać równania regresji zgodnie z uwagami zawartymi pomiędzy (W43) a (W44). Równanie (W42) jest treścią równania regresji pierwszego rodzaju. Jego współczynniki musimy oszacować na podstawie pobranej próbki odwołując się do MNW. Równanie regresji z oszacowanymi współczynnikami nazywamy *równaniem regresji drugiego rodzaju*. Funkcja wiarygodności dla analizy regresji Poissona ma ogólną postać (W46) [4, 1].

D1.6.1 Indeksowanie grup w przykładzie

Zgodnie z już wcześniej wprowadzonymi oznaczeniami, ponieważ mamy dwie populacje „Miast" i „Wsi", liczba generalnych populacji N=2 skąd, ze względu na poniżej wprowadzone kodowanie zmiennych ukrytych (tzn. kierunkowych), przyjmujemy j=0,1. Natomiast ze względu na występowanie jednego czynnika „wieku" samochodu X=X_1 , indeks i=k=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.

Zbudowanie modelu regresji Poissona dla powyższej sytuacji oznacza opisanie oczekiwanej liczby przypadków awarii silnika, $E(Y_{ij})$, poprzez wprowadzone do modelu zmienne objaśniające. Liczba zliczeń Y_{ij} jest zmienną losową Poissona (teoretycznie zmienną losową dwumianową) z wartością oczekiwaną równą $\mu_{ij} = \ell_{ij} r_{ij}$. Równanie to, dla określonej postaci

zależności r_{ij} od czynników, wyraża treść funkcji regresji pierwszego rodzaju, tzn. postulowaną jej postać w całej generalnej populacji[12].

Analiza regresji Poissona ma ustalić, czy widoczny „na oko" 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).

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 zgodnie z (W42) [4] następująco:

$$E(Y_{ij}) = \mu_{ij} = \ell_{ij} r_{ij} \quad , \qquad i = 1, 2, \ldots, 8 ; \quad j = 0, 1 . \tag{D6}$$

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

$$U_k = \begin{cases} 1 & \text{jeśli} \quad k = i , \qquad \text{gdzie} \quad i = 1, 2, \ldots, 7 \\ 0 & \text{w przeciwnym wypadku} \end{cases} \tag{D7}$$

$$M = \begin{cases} 1 & \text{jeśli} \quad j = 1 \qquad \text{(Wsie)} \\ 0 & \text{jeśli} \quad j = 0 \qquad \text{(Miasta)} \quad . \end{cases} \tag{D8}$$

[12] We wstępnych rozważaniach indeksowaliśmy grupy jednym indeksem n. Np. dla grupy n, gdzie $n = 1,2,\ldots,N$, zmienną obserwowanej ilości defektów oznaczaliśmy przez Y_n, a całkowitą wielkość zakumulowanego czasu dla wszystkich samochodów w n-tej podgrupie przez ℓ_n. Równanie regresji miało wtedy postać (W42) $\mu_n \equiv E(Y_n) = \ell_n r(x_n, \beta)$, $n = 1,2,\ldots,N$.

Podstawowa dla wielu analiz regresji Poissona, logarytmiczna postać funkcji ryzyka [6], która pojawi się w (D6) i korzystająca z kodowania (D7) oraz (D8) ma w przypadku **bez interakcji** następującą postać:

Model 1: $$\ln r_{ij} = \alpha + \sum_{k=1}^{7} \alpha_k U_k + \beta M \quad . \qquad \text{(D9)}$$

Korzystając z kodowania (D7) i (D8), możemy w powyższym „Modelu 1" ryzyka, wyrazić r_{ij} poprzez parametry α_i i β w następujący sposób:

$$\ln r_{i0} = \alpha + \alpha_i \quad \text{oraz} \quad \ln r_{i1} = \alpha + \alpha_i + \beta \ , \quad i = 1, 2, ..., 7, \qquad \text{(D10)}$$

oraz

$$\ln r_{80} = \alpha \quad \text{oraz} \quad \ln r_{81} = \alpha + \beta \ , \quad \text{dla } i = 8 \ , \qquad \text{(D11)}$$

co wynikało z tego, że $U_k = 1$ dla $k = i = 1, 2, ..., 7$ 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 (D10) i (D11) 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.

D1.7 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 (D10) i (D11) otrzymujemy:

$$\ln r_{i1} - \ln r_{i0} = (\alpha + \alpha_i + \beta - \alpha - \alpha_i) = \beta \quad , \qquad i = 1, 2, ..., 7 \qquad \text{(D12)}$$

oraz

$$\ln r_{81} - \ln r_{80} = (\alpha + \beta - \alpha) = \beta \ , \qquad i = 8 \ . \qquad \text{(D13)}$$

Korzystając z (D12) oraz (D13) widzimy, że **ryzyko względne** (D4) dla modelu (D9) nie zawierającego interakcji jest równe:

$$R_{Wi} = \frac{r_{i1}}{r_{i0}} = \exp\left[\ln\left(\frac{r_{i1}}{r_{i0}} \right) \right] = \exp[\ln r_{i1} - \ln r_{i0}] = \exp[\beta] = e^{\beta} \quad , \qquad i = 1, 2, ..., 8. \qquad \text{(D14)}$$

Powyższy model pozwala na estymację wskaźnika ryzyka względnego dla każdej kategorii wiekowej. Czynimy to stosując MNW do estymacji współczynnika kierunkowego β stojącego obok zmiennej M i w ten sposób dopasowując model do danych, a następnie licząc eksponentę tego estymatora.

Estymator ogólnego ryzyka względnego: Ponieważ estymowany wskaźnik ryzyka względnego e^{β} jest niezależny od i (tzn. od kategorii wiekowej), zatem możemy interpretować $\hat{r}_{Wi} = e^{\hat{\beta}}$ jako *estymator ogólnego ryzyka względnego* R_{Wi}, dopasowanego do wieku, gdzie $\hat{\beta}$ jest estymatorem MNW parametru β.

Wniosek *o postaci ryzyka względnego w modelu bez interakcji*: **Dla modelu** (D9) **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 k - czynników: X_1, X_2, ..., X_k. 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.

D1.8 Macierz kowariancji i obserwowana informacja Fishera

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

$$\hat{\mathrm{V}}(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,...,) = \begin{bmatrix} \hat{\sigma}^2(\hat{\beta}_0) & \hat{Cov}(\hat{\beta}_0,\hat{\beta}_1) & \hat{Cov}(\hat{\beta}_0,\hat{\beta}_2) \\ \hat{Cov}(\hat{\beta}_0,\hat{\beta}_1) & \hat{\sigma}^2(\hat{\beta}_1) & \hat{Cov}(\hat{\beta}_1,\hat{\beta}_2) \\ \hat{Cov}(\hat{\beta}_0,\hat{\beta}_2) & \hat{Cov}(\hat{\beta}_1,\hat{\beta}_2) & \hat{\sigma}^2(\hat{\beta}_2) \\ & & \ddots \end{bmatrix} := \mathbf{iF}^{-1} \quad , \qquad \text{(D15)}$$

oraz po drugie, 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].

D1.9 Statystyczne kryterium doboru modelu

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 (W56) oraz dewiację (W51), jako jego szczególny typ. W Rozdziale W1.3.3.2 przekonaliśmy się, że modele mogą być porównywane poprzez obliczenie różnic pomiędzy parami dewiancji tych modeli.

„**Model podstawowy**” został omówiony w Rozdziale W1.3.2. Wiarygodność próby $\tilde{Y}$ przyjmuje dla modelu podstawowego postać (W37) a jej postać zmaksymalizowana jest określona zgodnie z (W41).

Powód konstrukcji modelu regresji: Powodem analizowania modelu regresji, a nie trwania przy modelu podstawowym, nie jest sama dokładność dopasowania (która nie może być lepsza niż w modelu podstawowym), lecz próba zrozumienia istoty opisywanego zjawiska oraz mniejsza liczba parametrów, co wpływa na zmniejszenie kosztów oszacowywania parametrów z określoną dokładnością [4].

D1.9.1 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 $\mu_n \equiv E(Y_n) = \ell_n r(x_n, \beta)$, $n = 1,2,...,N$, zawierający $k+1$ parametrów, uznamy za oszczędny, jeśli ma wartość zmaksymalizowaną wiarygodności prawie tak dużą, jak 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.

D1.10 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 (D6) oraz (D9). Jest więc to uprzednio wprowadzony Model 1:

$$E(Y_{ij}) = \mu_{ij} = \ell_{ij} r_{ij} \quad , \qquad i = 1, 2, \ldots, 8 ; \quad j = 0, 1 , \tag{D16}$$

gdzie:

Model 1: $$\ln r_{ij} = \alpha + \sum_{k=1}^{7} \alpha_k U_k + \beta M \quad . \tag{D17}$$

Zmienne U_k były „sztucznie” wprowadzonymi zmiennymi kierunkowymi (ukrytymi) (D7) wskazującymi wariant wiekowy i przyjmującymi wartości 0 lub 1, a zmienna kierunkowa M przyjmowała zgodnie z (D8) wartości 0 lub 1, wskazując odpowiednio obszar Miejski lub Wiejski.

Dla powyższego modelu ryzyko względne wynosi (D5):

$$R_{Wi} = \frac{r_{i1}}{r_{i0}} \quad , \tag{D18}$$

a zgodnie z (D14) jego postać redukuje się do:

$$R_{Wi} = e^{\beta} \quad , \tag{D19}$$

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 (W46) ma ona w próbce postać:

$$P\left(y \mid (\beta)\right) = \prod_{i=1}^{8} \left\{ \left[\frac{\left(\ell_{i0} r_{i0}\right)^{y_{i0}} e^{-\ell_{i0} r_{i0}}}{y_{i0}!} \right] \left[\frac{\left(\ell_{i1} r_{i1}\right)^{y_{i1}} e^{-\ell_{i1} r_{i1}}}{y_{i1}!} \right] \right\} \quad , \tag{D20}$$

gdzie zgodnie z (D10)-(D11) mamy $r_{i0} = \exp(\alpha + \alpha_i)$ i $r_{i1} = \exp(\alpha + \alpha_i + \beta)$ dla i = 1, ..., 7 , oraz $r_{80} = \exp(\alpha)$ i $r_{81} = \exp(\alpha + \beta)$ dla i = 8 .

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

$$\left\{\hat{\alpha}, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\alpha}_4, \hat{\alpha}_5, \hat{\alpha}_6, \hat{\alpha}_7, \hat{\beta}\right\} \tag{D21}$$

oraz oszacowaną 9×9 - wymiarową macierz kowariancji (D15).

D1.11 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) [12]. 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 (zarówno dla programów wczytujących dane jak i procedur do analizy modeli), podane zostały w **Uzupełnieniu 1** na końcu dodatku. Pełny ich wykaz oraz zastosowanie można znaleźć w pomocy pakietu SAS.

D1.11.1 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 od A do O znajdują się w Uzupełnieniu 1.

Zbiór danych:

data awaria;

input A Y L M U1 U2 U3 U4 U5 U6 U7 U1M U2M U3M U4M U5M U6M U7M O;

ln = log(L);

datalines;

```
1   1   172675   0 1 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
3   30  96216    0 0 0 1 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
5   102 72159    0 0 0 0 0 1 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
7   133 32185    0 0 0 0 0 0 0 1 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
1   4   181343   1 1 0 0 0 0 0 0 1 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
3   119 121374   1 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0
4   221 111353   1 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0
5   259 83004    1 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0
6   310 55932    1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0
7   226 29007    1 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0
8   65  7538     1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
;
run;
```

Określenie funkcji wiążacej (Link Function) **oraz czynnika przesunięcia** (Offset Variable): Funkcja wiążacą $g(\mu_n)$ wiąże wartość oczekiwaną warunkową $\mu_n \equiv E(Y_n)$ z kombinacją liniową zmiennych objaśniających ujętą w postaci funkcji regresji. W przypadku rozkładu Poissona funkcja $g(\mu_n) = \ln(\mu_n)$, a w przypadku rozkładu normalnego $g(\mu_n) = \mu_n$.

W omawianych raportach SAS jest ona skonstruowana jako funkcja $g(r(x_n, \beta))$, zatem dla rozkładu Poissona otrzymujemy $g(r(x_n, \beta)) = \ln(r(x_n, \beta))$, gdzie funkcja ryzyka $r(x_n, \beta)$ ma postać jak w (W44). Ponieważ zgodnie z (W42), μ_n oraz $r(x_n, \beta)$ różnią się czynnikiem (skumulowanego czasu badania) $\ell_n \equiv \mathrm{L}$, zatem w powyższym programie dla zbioru danych, pojawiła się komenda ln = log(L) służąca do wczytania niezbędnego wyrażenia $\ln(\ell_n)$.

Natomiast na skutek przypisania w *poniższym* programie zmiennej '*offset*' wartości $\ln \equiv \ln(\ell_n)$, wyrażenie to pojawi się w raporcie SAS jako tzw. zmienna przesunięcia (Offset Variable).

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* = awaria*;*
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 = M A / corrb

a program SAS odda raport identycznej postaci.
Uwaga: Dodatkowo zamieniono komendę wyznaczenia macierzy kowariancji estymatorów ‘covb’ na polecenie ‘corrb’ wyznaczenia ich macierzy korelacyjnej.

D1.11.2 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
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. (Przedział ufności Wald'a, patrz Uzupełnienie 2).

D1.11.3 Oszacowanie parametru i błąd standardowy oszacowania dla Modelu 1

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

$$\hat{\beta} = 0{,}8043 \tag{D22}$$

oraz *błąd standardowy* (*se*) tego oszacowania wyznaczony jako element macierzy (wariancji-) kowariancji (D15) [4]:

$$\hat{\sigma}_{\hat{\beta}} \equiv se(\hat{\beta}) = (0.002725)^{1/2} = 0{,}0522\ . \tag{D23}$$

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

$$\hat{r}_{Wi}^{Model1} = e^{\hat{\beta}} = e^{0{,}8043} = 2{,}23513\ . \tag{D24}$$

Natomiast 95%-owy wiarygodnościowy przedział ufności Wald'a (patrz Uzupełnienie 2) dla e^{β} [4], przy odwołaniu się do faktu, że dla dużej próbki estymator MNW ma *w przybliżeniu* rozkład normalny, ma postać:

$$\exp[\hat{\beta} \pm 1{,}96\ \hat{\sigma}_{\hat{\beta}}] = \exp[0{,}8043 \pm 1{,}96\ (0{,}0522)] = \exp(0{,}8043 \pm 0{,}1023)\ , \tag{D25}$$

lub

$$\left(e^{0.7020}, e^{0.9066}\right) = (2.01778;\ 2{,}47589)\ . \tag{D26}$$

D1.11.4 Test hipotezy zerowej z wykorzystaniem statystyki Wald'a

Dla dużej próbki, test hipotezy zerowej:

$$\mathrm{H}_0\colon\ \beta = 0 \tag{D27}$$

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

$$H_1: \beta \neq 0 \ , \tag{D28}$$

może być przeprowadzony z zastosowaniem statystyki Wald'a [4] (patrz Uzupełnienie 2):

$$U = \frac{\hat{\beta} - 0}{\hat{\sigma}_{\hat{\beta}}} \quad . \tag{D29}$$

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

Dla rozważanego przykładu wartość statystyki Wald'a wynosi:

$$U = \frac{0{,}8043 - 0}{0{,}0522} = 15{,}408 \ , \tag{D30}$$

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

$$p = \Pr\left(|U| \geq 15{,}408\right) < 0{,}0001 \ . \tag{D31}$$

D1.11.5 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 wiarygodnościowy 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.

D1.12 Charakter kowarianta „wiek” - interakcja czy zaburzenie

Głównym wpływem interesującym nas w analizie ryzyka jest zmienna „obszar”. W formule (D14) na ryzyko względne, zmienna wiek okazała się nawet nie występować. Jednak w wyprowadzeniu (D14) 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 Punktu 1, 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].

D1.12.1 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 , (D17) (porównaj (D9)), o człon interakcji, otrzymując:

Model 2: $\ln r_{ij} = \alpha + \sum_{k=1}^{7} \alpha_k U_k + \beta M + \sum_{k=1}^{7} \delta_k (MU_k)$, i =1, 2,..., 8; j = 0, 1 . (D32)

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:

$$H_0: \delta_1 = \delta_2 = \ldots = \delta_7 = 0 \ , \quad \text{(D33)}$$

z wykorzystaniem statystyki ilorazu wiarygodności (W51). 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 (W51) pozwala więc na porównanie Modelu 1 (bez interakcji) z Modelem 2, który zawiera siedem iloczynowych członów interakcji $M U_k$.

D1.12.2 Program SAS dla Modelu 2

Ponieważ w Modelu 2 chcemy uwzględnić również interakcję „wieku" i „obszaru", zatem po wczytaniu danych takich samych jak w Punkcie 1.11.1, należy przy korzystaniu z procedury GENMOD (Punkt 1.11.1) zmienić linię *model* na uwzględniający człony interakcji $M U_k$, k=1,2,...,7, wczytując program:

```
proc genmod data = awaria;
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;
```

D1.12.3 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	St. sw.	Ocena	Błąd standardowy	95% granice przedziału ufności Walda		Chi-kwadrat	Pr > chi kw..
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:

$$D\left(\hat{\beta}\right)^{Model\,2} = 0 \ , \tag{D34}$$

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 (W51) dla Modelu 1 i Modelu 2.

D1.12.4 Testowanie braku dopasowania w Modelu 1 w porównaniu z Modelem 2

Rozważmy hipotezę zerową (D33):

$$H_0: \delta_1 = \delta_2 = \ldots = \delta_7 = 0 \qquad \text{(D33')}$$

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

Okazuje się, że w rozważanym przypadku test statystyczny weryfikujący hipotezę zerową (D33), można by przeprowadzić zarówno wykorzystując statystykę ilorazu wiarygodności (co jest oczywiste), jak i dewiancję Modelu 1.
Istotnie, po pierwsze w Rozdziałach W1.3.3.1- W1.3.3.2 zauważyliśmy, że obie te statystyki mają w przybliżeniu rozkład chi-kwadrat [4]. Po drugie, zauważmy, że dewiancja dla Modelu 1 otrzymana w raporcie w D**1.11.2** przyjęła w próbce wartość:

$$D\left(\hat{\beta}_{(r)}\right)^{Model1} = 8{,}195 \ . \qquad \text{(D35)}$$

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

$$d.f. = [\textit{liczba zmiennych } (Y_{ij})] - [\textit{liczba parametrów w } \text{Modelu 1}] = N - (r+1)$$
$$= 16 - 9 = 7. \qquad \text{(D36)}$$

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

Z kolei statystyka testowa ilorazu wiarygodności (W57) dla hipotezy zerowej (D33) jest zgodnie z (W58) otrzymana przez odjęcie dewiancji dla Modelu 2 (która jest równa zero) od dewiancji dla Modelu 1, tzn.:

$$-2\ln\left[\frac{P\left(\tilde{\tilde{Y}} \mid \hat{\beta}_{(r)}\right)}{P\left(\tilde{\tilde{Y}} \mid \hat{\beta}\right)}\right] = D\left(\hat{\beta}_{(r)}\right)^{Model1} - D\left(\hat{\beta}\right)^{Model2} = 8.195 - 0 = 8.195 \quad , \qquad \text{(D37)}$$

zatem jej wartość w próbce jest równa $D\left(\hat{\beta}_{(r)}\right)^{Model1}$ jak w (D35).

Również liczba stopni swobody statystyki ilorazu wiarygodności (W51), równa [1]:

$$d.f. = [\textit{liczba parametrów w} \text{ Modelu 2}] - [\textit{liczba parametrów w} \text{ Modelu 1}]$$

$$= 16 - 9 = 7 \, , \qquad \text{(D38)}$$

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

Zbierzmy informacje zawarte we wzorach od (D35) do (D38). Wynika z nich, że skoro zarówno rozkład, jak i wartość liczbowa oraz liczba stopni swobody dewiancji Modelu 1, (D35), oraz log ilorazu wiarygodności, (D37), są takie same, zatem równoważnie można weryfikować hipotezę zerową (D33) korzystając ze statystyki (D37) bądź (D35) (por. Uwaga na końcu Rozdziału W1.3.3.2).

Przyjmijmy więc, w tym przypadku, dewiancję $D\left(\hat{\beta}_{(r)}\right)^{Model1}$ dla Modelu 1 jako statystykę testową hipotezy (D33). Korzystając z (D35) oraz (D36) otrzymujemy, wykonując pomocnicze rachunki na przykład w arkuszu kalkulacyjnym Excel, że empiryczny poziom istotności wynosi:

$$p = \Pr\left(\chi_7^2 \geq D\left(\hat{\beta}_{(r)}\right)^{Model1} = 8{,}195\right) = 0.3157 \, . \qquad \text{(D39)}$$

D1.12.4.1 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.

D1.12.5 Analiza „wieku" jako zaburzenia czynnika głównego

Rozważenie Punktu 2 w Rozdziale D1.12 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].

D1.12.5.1 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* różnić się będzie **znacząco** od wartości $\hat{\beta} = 0{,}8043$, (D22), 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*.

D1.12.5.2 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 (D40)

Zadanie dla Modelu 3. Napisać program korzystający z procedury GENMOD dla Modelu 3, a następnie wykorzystując dane podane w Punkcie 1.11.1 uruchomić go, otrzymując poniższy raport SAS.

D1.12.5.3 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.

D1.12.5.4 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:

$$\hat{r}_W^{Model3} = e^{\hat{\beta}} = e^{0,7431} = 2,1024 \ . \qquad \text{(D41)}$$

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.*

D1.12.5.5 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 (W58), 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.

D1.13 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 . (D42)

D1.13.1 Dane i program SAS dla Modelu 0

Aby przeprowadzić analizę z użyciem procedury GENMOD została w danych podanych w Rozdziale D1.11.1 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 / covb

lub

model Y = O / pred covb

D1.13.2 Raport SAS dla Modelu 0

Po wczytaniu danych zawartych w Rozdziale D1.11.1 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.

D1.13.3 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, (D42), w stosunku do dewiancji Modelu 1, (D17).

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, (D42), oraz Modelu 3, (D40), jest również statystycznie istotna, znajdując wartość odpowiedniego empirycznego poziomu istotności.

D1.14 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	

D1.14.1 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 1: 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 'awaria' 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;
corrb wskazuje na wyliczenie macierzy korelacyjnej 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.

Uzupełnienie 2: Błąd statystyczny i statystyka Wald'a

Wiarygodnościowe przedziały ufności wspomniane w Rozdziale W1.2.1 są uzupełnieniem analizy MNW pozwalającym na szybkie określenie niepewności oszacowania skalarnego parametru θ. Posługiwanie się nimi jest prostrze niż samą funkcją wiarygodności. Istotnym pojęciem przy ich konstrukcji jest obserwowana informacja Fishera $\mathbf{iF}\left(\hat{\theta}\right)$ wprowadzona w Rozdziale W1.2.1 jako czynnik po prawej stronie wyrażenia (W25):

$$\ln \frac{P(y \mid \theta)}{P\left(y \mid \hat{\theta}\right)} \approx -\frac{1}{2}\mathbf{iF}\left(\hat{\theta}\right)\left(\hat{\theta}-\theta\right)^2. \qquad \text{(U2.1)}$$

Wyrażenie to jest w przybliżeniu słuszne dla *rozkładów regularnych*, dla których funkcjonuje przybliżenie kwadratowe dla log ilorazu funkcji wiarygodności. Więcej na temat $\mathbf{iF}\left(\hat{\theta}\right)$ można znaleźć w [1,6].

W przypadku gdy pierwotna zmienna losowa Y ma rozkład normalny $N(\theta,\sigma^2)$, wtedy związek (W25) staje się dokładny, a z (W16) widać, że $\mathbf{iF}\left(\hat{\theta}\right)=\frac{N}{\sigma^2}$. Ponadto, dla rozkładu normalnego można rozkładem χ^2_1 dokonać dokładnego wyskalowania ilorazu funkcji wiarygodności, a zatem i wartości parametru obcięcia $c = e^{-\frac{1}{2}\chi^2_{1,(1-\alpha)}}$, zgodnie z (W22).

Błąd standardowy: Obserwowana informacja Fishera definiuje tzw. *błąd standardowy* estymatora $\hat{\theta}$ jako równy:

$$\hat{\sigma}_{\hat{\theta}} \equiv se\left(\hat{\theta}\right)=\left(\mathbf{iF}\left(\hat{\theta}\right)\right)^{-1/2}, \qquad \text{(U2.2)}$$

który w przypadku rozkładu normalnego wynosi:

$$\hat{\sigma}_{\hat{\theta}} = \sigma/\sqrt{N}\,. \qquad \text{(U2.3)}$$

Jeśli przyjąć, że iloraz wiarygodności jest równy wartości granicznej $P(y \mid \theta)/P\left(y \mid \hat{\theta}\right)=c$, wtedy dla przedziału wiarygodności (W24), $\left\{\theta, P\left(\tilde{Y} \mid \theta\right)/P\left(\tilde{Y} \mid \hat{\theta}\right) > c\right\}$ otrzymujemy z (W25) graniczną wartość parametru $\theta \approx \hat{\theta} \pm \sqrt{-2\ln c}\,\left(\mathbf{iF}\left(\hat{\theta}\right)\right)^{-1/2}$. Stąd przybliżona postać[13] przedziału wiarygodności jest następująca $\left(\hat{\theta}-\sqrt{-2\ln c}\;\hat{\sigma}_{\hat{\theta}},\;\; \hat{\theta}+\sqrt{-2\ln c}\;\hat{\sigma}_{\hat{\theta}}\right)$, co w

[13] Dla rozkładu normalnego jest to postać dokładna.

przypadku rozkładu normalnego daje dokładną postać $(1-\alpha)\cdot 100\%$-wego przedziału wiarygodności (CI):

$$\left(\hat{\theta} - \sqrt{\chi^2_{1,(1-\alpha)}}\ \hat{\sigma}_{\hat{\theta}}\ ,\quad \hat{\theta} + \sqrt{\chi^2_{1,(1-\alpha)}}\ \hat{\sigma}_{\hat{\theta}}\right) \qquad \text{(U2.4)}$$

Na przykład, gdy $(1-\alpha) = 0{,}95$, wtedy 95%-wy dokładny przedział wiarygodności wynosi:

$$\hat{\theta} \pm \sqrt{1{,}96}\ \hat{\sigma}_{\hat{\theta}} \qquad \text{(U2.5)}$$

W przypadku reguralnym przedział (U2.4) określa przybliżoną postać $(1-\alpha)\cdot 100\%$-wego przedziału wiarygodności.

Test Wald'a dla hipotezy zerowej o skalarnym parametrze θ: Zweryfikujmy hipotezę zerową $H_0 : \theta = \theta_0$ wobec hipotezy alternatywnej $H_1 : \theta \neq \theta_0$. W celu przeprowadzenia testu statystycznego wprowadźmy tzw. statystykę Wald'a.

Statystyka Wald'a ma postać:

$$U = \frac{\hat{\theta} - \theta_0}{\hat{\sigma}_{\hat{\theta}}} \quad , \qquad \text{(U2.6)}$$

gdzie wartość *u* zmiennej *U* jest wyznaczona na podstawie obserwacji i dla estymatora $\hat{\theta}$ MNW parametru θ. Z porównania (U2.1) oraz (U2.6), widać, że duża wartość $|u|$ statystyki $|U|$ jest związana z małą wiarygodnością modelu dla $H_0 : \theta = \theta_0$.

Przykład: Niech na podstawie obserwacji wartość $|u| = 3$. Zakładając regularność modelu, otrzymujemy z (U2.1) oraz (U2.6) i przy wartości granicznej $\frac{P(y \mid \theta)}{P(y \mid \hat{\theta})} = c$ ilorazu wiarygodności, związek pomiędzy parametrem *c* a wartością *u*:

$$c = \exp(-\frac{u^2}{2}) \ . \qquad \text{(U2.7)}$$

Dla rozważanego przykładu z (U2.7) otrzymujemy $c = \exp(-\frac{u^2}{2})\Big|_{|u|=3} = \exp(-4{,}5) = 0{,}011$. Zatem zgodnie z (W27) wartość empirycznego poziomu istotności wynosi $p = P(\chi_1^2 > -2\ln c) = P(\chi_1^2 > 9) = 0{,}0027$. Oznacza to, że na każdym poziomie istotności

$\alpha \geq p = 0{,}0027$, np. $\alpha = 0{,}01$, odrzucamy hipotezę zerową $H_0 : \theta = \theta_0$ na rzecz hipotezy alternatywnej.

W Rozdziałach **D1.11.3 i D1.11.4** zastosowano statystykę Wald'a do estymacji i weryfikacji hipotezy zerowej dotyczącej parametru β Modelu 1, (D17).

Zakończenie

Przedmiotem Dodatku 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. Niemniej estymacja MNW jest szczególnie przydatna w estymacji współczynników regresji w modelach nieliniowych, takich jak model regresji logistycznej czy nieliniowy 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] D. Mroziakiewicz, *Analiza regresji Poissona z estymatorami metody największej wiarygodności z wykorzystaniem programu statystycznego SAS*, Praca licencjacka, Inst. Fizyki, Uniwersytet Śląski, (Rybnik), (2006).

[8] M. Czerwik, *Wykorzystanie programu SAS jako narzędzia do analizy współzależności zmiennych metoda regresji*, Praca licencjacka, Uniwersytet Śląski, Inst. Fizyki, Jastrzębie Zdrój, (2004).

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

[10] M. Biesiada, *Statystyka w ujęciu Bayesowskim*, Skrypt dla studentów ekonofizyki, Uniwersytet Śląski, Instytut Fizyki, (2011).

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

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