Analiza danych ankietowych
Lista 6-7

1. Używając metody największej wiarogodności wyznaczyć estymatory parametrów modelu liniowej regresji wielokrotnej.

2. Wyznaczyć postać statystyki ilorazu wiarogodności
$$G^2 = -2\log \Lambda,$$
gdzie
$$\Lambda = \frac{\text{maksimum funkcji wiarogodności dla modelu przy } H_0}{\text{maksimum funkcji wiarogodności dla modelu przy } H_0 \cup H_a}$$
dla testowania niezależności w tablicach kontyngencji 2×2. Z iloma stopniami swobody statystyka G^2 ma rozkład χ^2 przy H_0?

3. Wyznaczyć statystykę dewiancji (odchylenia) $D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$ dla modeli Poissona *GLM*. Pokazać, że dla tabel kontyngencji 2×2 redukuje się ona do statystyki G^2.

4. Dla n niezależnych obserwacji z rozkładu Poissona pokazać, że w przypadku metody wynikowej Fishera (*Fisher scoring method*) otrzymano $\mu^{(t+1)} = \overline{y}$, dla wszystkich $t > 0$. Co otrzymamy w przypadku algorytmu Newtona-Raphsona?

5. Napisać program komputerowy używając algorytmu Newtona-Raphsona do maksymalizacji wiarogodności próby dwumianowej. Dla $\hat{\pi} = 0.3$ na podstawie $n = 10$, wydrukować wyniki pierwszych sześciu iteracji przy uruchomieniu z wartością początkową $\pi^{(0)}$ równą a) 0.1, b) 0.2, ... i) 0.9. Podsumować efekty wartości początkowej na prędkość zbieżności. Co się stanie, jeśli punktem startowym będzie 0 lub 1?

6. Rozważyć klasę modeli binarnych
$$\pi(x) = \Phi(\alpha + \beta x), \tag{1}$$
gdzie Φ jest dystrybuantą z parametrem przesunięcia 0 i skali 1, natomiast φ jest gęstością tego rozkładu symetryczną względem zera.

   a) Uzasadnić, że największe nachylenie krzywa regresji (1) ma dla $\pi(x) = 0.5$.

   b) Wyznaczyć tempo zmiany $\pi(x)$, jeśli $\pi(x) = 0.5$.

   c) Pokazać, że dla funkcji wiążących: *link=logit* oraz *link=probit* tempo zmiany $\pi(x)$ (dla $\pi(x) = 0.5$) wynosi odpowiednio 0.25β oraz $\beta/\sqrt{2\pi}$.

TABLE 4.7 SAS Output for Problem 4.7

Criterion	DF	Value
Deviance	171	560.8664
Pearson Chi-Square	171	535.8957
Log Likelihood		71.9524

Parameter	Estimate	Std Error	Wald 95% Conf Limits		Chi-Sq	Pr > ChiSq
Intercept	−0.4284	0.1789	−0.7791	−0.0777	5.73	0.0167
weight	0.5893	0.0650	0.4619	0.7167	82.15	<.0001

7. Dane z Tabeli 4.3 (Agresti (2002), str. 127) dotyczą badania gniazd krabów (skrzypłoczy). W Tabeli 4.7 pokazano wyniki pakietu SAS dla dopasowania log-liniowego modelu Poissona, gdy zmienna objaśniająca $X =$ waga i zmienna objaśniana $Y =$ liczba satelitów.

   a) Oszacować EY dla samicy kraba o średniej wadze 2,44 kg.

   b) Zbadać testem Walda, czy Y jest niezależne od X.

   c) Czy można przeprowadzić test ilorazu wiarogodności tej hipotezy? Jeśli nie, co jeszcze jest potrzebne?

   d) Wyznaczyć przedziały ufności parametrów tego modelu.

8. Napisać równania wiarogodności i macierz kowariancji dla log-liniowego modelu Poissona.

9. Wyznaczyć Hesjan (macierz drugich pochodnych) oraz uzasadnić, że oba algorytmy wynikowy Fishera (*Fisher scoring*) i Newtona-Raphsona są identyczne dla log-liniowego modelu Poissona.

10. Napisać program komputerowy do wyznaczania estymatorów parametrów log-liniowego modelu Poissona używając algorytmu Newtona-Raphsona.

11. Dla przykładu z zadania 7 wyznaczyć i narysować 95% przedziały ufności dla średniej liczby satelitów w zależności od zmieniającej się wagi pancerza samicy kraba.

12. Przykład dotyczący badania gniazd skrzypłoczy (dane z Tabeli 4.3, str. 127 Agresti (2002)). Binarna zmienna objaśniana w tym przykładzie to: $Y = 1$, jeśli samica kraba ma co najmniej jednego satelitę oraz $Y = 0$, jeśli samica kraba nie ma żadnego samca w pobliżu gniazda. Rozważyć model (Agresti (2002), str. 188)

$$logit(\pi) = \alpha + \beta_1 c_1 + \beta_2 c_2 + \beta_3 c_3 + \beta_4 x, \tag{2}$$

   gdzie c_1, c_2 i c_3 to zmienne zero-jedynkowe (*dummy variables*) odpowiednio dla kolorów pancerzy *„medium-light”*, *„medium”*, *„medium-dark”*. Ponadto $\pi = P(Y = 1)$.

   a) Dopasować model za pomocą $x =$ waga pancerza samicy (w kilogramach). Zinterpretować efekty wagi i kolorów. Wyznaczyć $logit(\hat{\pi})$ dla ciemnego (*„dark”*) koloru pancerza. Ile razy wzrośnie estymowana szansa pojawienia się satelitów jeśli waga pancerza zwiększy się o 1kg?

   b) Czy model umożliwiający interakcję zapewnia lepsze dopasowanie?

   c) Skonstruować przedział ufności dla różnicy między parametrami nachylenia dla średnio-jasnych (*„medium-light”*) i ciemnych (*„dark”*) krabów.

   d) Używając poniższego modelu traktującego kolor jako zmienną ilościową

$$logit(\pi) = \alpha + \beta_1 c + \beta_2 x, \tag{3}$$

   gdzie $c = \{1, 2, 3, 4\}$ dla poszczególnych kategorii zmiennej kolor, powtórzyć analizy (patrz Agresti (2002), str. 190).