Uogólnione modele liniowe

Laboratorium nr 9

9.1 Rozważamy zbiór gator.data, z danymi o 219 aligatorach schwytanych w czterech jeziorach Florydy (odpowiednio: Hancock, Oklawaha, Trafford, George). Zmienną (nominalną) odpowiedzi jest food (pięć kategorii pożywienia znajdowanego w żołądkach krokodyli, odpowiednio: ryby, bezkręgowce, gady, ptaki, inne). Zmienna size ma dwie wartości, odpowiadające kategoriom $\leq 2.3m$ i $> 2.3m$. Zmienna gender ma dwie wartości (1=male, 2=female). Celem zadania jest zbadanie wpływu zmiennych lake (L), gender (G) i Size (S) na typ pożywienia (food,F).

(a) Za pomocą funkcji multinom z biblioteki nnet, która dopasowuje modele logitowe dla odpowiedzi nominalnych, dopasować modele:

i. minimalny,
ii. F∼G, F∼S, F∼L
iii. analogicznie modele uwzględniające addytywny wpływ par zmiennych G,S,L
iv. model F∼G+S+L
v. model wysycony F∼ $G * S * L$ (kodowanie

```
fitS<-multinom(food~lake*size*gender,data=...)
```

Jak wybiera się kategorię referencyjną w multinom?

(b) Dla każdego z modeli zbadać jakość dopasowania obliczając

```
deviance(model)-deviance(model wysycony)
```

(c) Przeprowadzić analogiczne analizy po zagregowaniu danych ze względu na G.

(d) Dla danych zagregowanych ze względu na płeć i modelu F∼L+S, obliczyć wartości dopasowane (fitted values) i porównać je z wartościami obserwowanymi.

(e) Na podstawie powyższego punktu obliczyć wartość statystyki X^2 i porównać ją z odpowiednią różnicą dewiacji.

(f) Dla danych zagregowanych ze względu na płeć i modelu F∼L+S, pisząc

```
library(MASS) # potrzeba funkcji vcov
summary(nazwa modelu, cor = F)
```

oszacować wpływ jeziora i rozmiaru aligatora na szanse tego, że wybierze on inne niż ryby zasadnicze źródła pożywienia. Sprawdzić w szczególności, że równanie predykcyjne dla logarytmu szans wyboru bezkręgowców zamiast ryb to

$$\log\left(\hat{\pi}_{\text{bezkreg}}/\hat{\pi}_{\text{ryby}}\right) = -1.55 + 1.46s - 1.66z_H + 0.94z_O + 1.12z_T, \tag{1}$$

gdzie $s = 1$, gdy rozmiar jest ≤ 2.3 i $s = 0$ w przeciwnym przypadku, z_H jest zmienną indykatorową (dummy variable) dla jeziora Hancock (tzn. $z_H = 1$, gdy aligator pochodzi z Hancock i $z_H = 0$ w p.p.), z_T i z_O to zmienne indykatorowe dla jezior Trafford i Oklawaha.

(g) Na podstawie poprzedniego punktu: wyestymować prawdopodobieństwo tego, że duży aligator z jeziora Hancock wybierze bezkręgowce jako główne źródło swojego pożywienia.

(h) Eksperymentalnie ustalić (np. na podstawie modelu F∼L+S) wpływ wyboru kategorii referencyjnej na estymatory $\widehat{\beta}$ i $\widehat{\pi}$.