

Uogólnione modele liniowe

Laboratorium nr 10

10.1 (Model proporcjonalnych szans) Dane zawarte w pliku `impair.data` pochodzą z badania zdrowia psychicznego losowej próby dorosłych mieszkańców hrabstwa Alachua w stanie Floryda (USA). Zmienna *mental* (zmienna ze skali porządkowej) określa stan zdrowia psychicznego (kategorie: dobry(1), łagodny (2), umiarkowany (3), zaburzony (4)). Zmienne objaśniające to *events* (tzw. life events index – zbiorcza miara liczby i intensywności przeżyć w rodzaju urodzenie dziecka, nowa praca, rozwód, śmierć członka rodziny - wszystko w ciągu ostatnich trzech lat) i *ses* (status socjoekonomiczny – tutaj mierzony w sposób binarny, 1=wysoki, 0=niski).

- (a) Przekształcić zmienną *mental* do czynnika uporządkowanego (*ordered factor*), z kategoriami jak powyżej. Dopasować logitowy model proporcjonalnych szans za pomocą procedury `polr` (ang. *proportional odds logistic regression*) z MASS:

```
g=polr(mental~ses+events,data=...).
```

- (b) Obliczyć prawdopodobieństwo bycia w kategorii *mental*=1(=stan dobry) przy *ses*=0 i wartości zmiennej *events*=4.275.
- (c) Narysować wykres wyestymowanych wartości $\mathbb{P}(Y > 2)$ (*Y* oznacza zmienną odpowiedzi) jako funkcji *events* na dwóch poziomach SES: 0 i 1 (jeden wykres).
- (d) Zauważyć, że szanse bycia w kategorii *mental*=1 wzrastają około trzykrotnie przy wysokim statusie socjoekonomicznym (*ses*=1) w porównaniu z niskim, tj. *ses*=0 (przy dowolnej ustalonej wartości kategorii *events*) – taka sama odpowiedź dla szans bycia poniżej dowolnego poziomu zdrowia psychicznego.
- (e) W celu zbadania wpływu zmiennej *ses* na odpowiedź, obliczyć, ile wynosi $\hat{\mathbb{P}}(Y = 1)$ dla *ses*=1 i zmiennej *events* przyjmującej swoją średnią wartość (i porównać to z wartością otrzymaną w punkcie (1b)).
- (f) W celu zbadania wpływu zmiennej *events* na odpowiedź, obliczyć, jak zmienia się $\hat{\mathbb{P}}(Y = 1)$ dla przejścia od dolnego do górnego kwartyla zmiennej *events* (osobno dla *ses*=0 i *ses*=1).

10.2 (Modele logliniowe dla tablicy dwudzielczej) Rozpatrzmy ponownie zbior `gator.data`. Celem zadania jest zbadanie zależności między zmienną *lake* a zmienna *food*.

- (a) Zagregować dane do tablicy kontyngencji (*lake,food*).
- (b) Celem jest testowanie hipotezy $p_{i,j} = p_i * p_j$, czyli

$$\log(\text{oczekiwana licznosc w klatce } (i,j)) = \log n + \log p_i + \log p_j$$

Przetestować powyższą hipotezę przez dopasowanie stosownego modelu poissonowskiego.

- (c) Dla każdego jeziora obliczyć frakcję aligatorów z niego pochodzących i porównać ją z estymowanym prawdopodobieństwem, że aligator pochodzi z tego jeziora.

10.3 Komendy

```
library(faraway)
data(femsmoke)
```

dają dostęp do danych związanych z następującym badaniem. W latach 1972-74 badano pod różnymi kątami grupę kobiet, które m.in. podzielono na palące/niepalące i sklasyfikowano pod względem wieku. Po 20 latach sprawdzano, które z badanych kobiet żyją.

- (a) Przeczytać opis danych ze zbioru `femsmoke`.
- (b) Stworzyć tablicę kontyngencji dla zmiennych: paląca/niepaląca i żyje/nie żyje. Wyliczyć proporcje osób żyjących i nieżyjących dla palaczek i niepalących. Zauważyć, że 76% palaczek przeżyło 20 lat, podczas gdy wśród niepalących analogiczny odsetek to jedynie 69%.
- (c) Powtórzyć wyliczenia z poprzedniego punktu dla kobiet z każdej grupy wiekowej z osobna.
- (d) Stworzyć tabelkę z proporcjami paląca/niepaląca w każdej grupie wiekowej. Wyjaśnić na jej podstawie paradoks z punktu (b).
- (e) Obliczyć stosunki szans na podstawie tabeli z punktu (b).
- (f) Sprawdzić niezależność zmiennych *smoker*, *dead* i *age* poprzez zbadanie jakości dopasowania odpowiedniego modelu poissonowskiego.
- (g) Sprawdzić niezależność *age* i pary (*smoker,dead*) poprzez zbadanie jakości dopasowania odpowiedniego modelu poissonowskiego.
- (h) Ustalić, jakiemu rodzajowi niezależności odpowiada model:

```
glm(y ~ smoker*age + age*dead, femsmoke, family=poisson)
```

Zbadać jego dopasowanie. Czy jest to model wysycony?

(i) Rozważyć model

```
glm(y ~ (smoker+age+dead)^2, femsmoke, family=poisson)
```

Czy jest to model wysycony? Dla każdej grupy wiekowej obliczyć wartości dopasowane par (smoker,dead) w tym modelu i wyznaczyć stosunki szans w każdej grupie wiekowej opierając się na wyliczonych wartościach dopasowanych (zauważyć równość wyliczonych stosunków szans w każdej grupie wiekowej). Powtórzyć obliczenia dla pozostałych (znaczących w tym kontekście) kombinacji zmiennych. Zinterpretować rozważany model.

(j) Rozważyć model

```
glm(y ~ smoker*age*dead, femsmoke, family=poisson)
```

Czy jest to model wysycony? Sprawdzić, czy można z niego usunąć interakcję trzeciego rzędu.