Zadanie z analizy log-liniowej

Agnieszka Wrzos

Zadanie

W 1996 roku w USA przeprowadzono badania wśród licealistów na temat stosowania następujących używek: alkohol, papierosy i marihuana. Poniżej znajduje się tabela kontygencji w formie płaskiej opisująca wyniki:

```
##   cigarette marijuana alcohol freq
## 1       yes       yes     yes  911
## 2        no       yes     yes   44
## 3       yes        no     yes  538
## 4        no        no     yes  456
## 5       yes       yes      no    3
## 6        no       yes      no    2
## 7       yes        no      no   43
## 8        no        no      no  279
```

Na podstawie tych danych dopasuj właściwy model log-liniowy i zinterpretuj wyniki.

Rozwiązanie

Dane

Zmienna `dane` zawiera tabelę z 2276 obserwacjami dot. odpowiedzi na temat używek.

```
dane <- data.frame(cigarette="yes", marijuana="yes", alcohol="yes")
f=910
nwiersz = data.frame(cigarette="yes", marijuana="yes", alcohol="yes")
for (k in (1:f)){
  dane <- rbind(dane, nwiersz)
}
x=c(2:8)
for (i in x){
  f=tabela$freq[i]
  nwiersz = data.frame(cigarette=tabela$cig[i], marijuana=tabela$mari[i], alcohol=tabela$alco[i])
  for (j in (1:f)){
    dane <- rbind(dane, nwiersz)
  }
}
```

```
tab <- table(dane)
tab
```

```
## , , alcohol = no
##
##          marijuana
## cigarette  no yes
```

```
##       no  279   2
##       yes  43   3
##
## , , alcohol = yes
##
##          marijuana
## cigarette  no yes
##       no  456  44
##       yes 538 911
```

```
ftab <- ftable(dane)
ftab
```

```
##                     alcohol  no yes
## cigarette marijuana
## no        no                279 456
##           yes                 2  44
## yes       no                 43 538
##           yes                 3 911
```

Tabela kontygencji przedstawia wszystkie kombinacje odpowiedzi i porównanie ich występowania. Widać, że największa liczba studentów miała styczność z trzema wymienionymi używkami. Można też wywnioskować, że osoby które nie próbowały alkoholu nie sięgają raczej także po pozostałe używki.

Budowa modeli

Stosując budowę hierarchiczną stworzę modele.

```
mod0 <- loglm(~cigarette+alcohol+marijuana,data=tab)
mod0
```

```
## Call:
## loglm(formula = ~cigarette + alcohol + marijuana, data = tab)
##
## Statistics:
##                       X^2 df P(> X^2)
## Likelihood Ratio 1286.020  4        0
## Pearson          1411.386  4        0
```

Model nie jest dobrze dopasowany($p<0.1$). Dodaję interakcje rzędu 2.

```
mod2 <- update(mod0, ~.^2)
mod2
```

```
## Call:
## loglm(formula = ~cigarette + alcohol + marijuana + cigarette:alcohol +
##     cigarette:marijuana + alcohol:marijuana, data = tab)
##
## Statistics:
##                        X^2 df  P(> X^2)
## Likelihood Ratio 0.3739859  1 0.5408396
## Pearson          0.4011039  1 0.5265197
```

Nie ma podstaw do odrzucenia hipotezy o dopasowaniu modelu, więc interakcje wyższego rzędu nie są potrzebne. Następnym krokiem będzie sprawdzenie czy wszystkie interkacje drugiego rzedu są niezbędne.

```
add1(mod0, test = "Chisq", scope = mod2)
```

```
## Single term additions
##
## Model:
## ~cigarette + alcohol + marijuana
##                     Df     AIC    LRT  Pr(>Chi)
## <none>                 1294.02
## cigarette:alcohol    1  853.83 442.19 < 2.2e-16 ***
## cigarette:marijuana  1  544.21 751.81 < 2.2e-16 ***
## alcohol:marijuana    1  949.56 346.46 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Nie trzeba nic zmieniać, bo uproszczenie modelu wpłynie negatywnie na jego dopasowanie. Poniżej upewniam się porównując dwa utworzone modele.

```
anova(mod0, mod2)
```

```
## LR tests for hierarchical log-linear models
##
## Model 1:
##  ~cigarette + alcohol + marijuana
## Model 2:
##  ~cigarette + alcohol + marijuana + cigarette:alcohol + cigarette:marijuana + alcohol:marijuana
##
##               Deviance df   Delta(Dev) Delta(df) P(> Delta(Dev)
## Model 1   1286.0199544  4
## Model 2      0.3739859  1 1285.6459685         3        0.00000
## Saturated    0.0000000  0    0.3739859         1        0.54084
```

Model mod2 osiąga satysfakcjonujący poziom dopasowania.

Postać addytywna i multiplikatywna modelu

Postać addytywna modelu:

$$log(\hat{n}_{ijk}) = \lambda + \lambda_i^C + \lambda_j^A + \lambda_k^M + \lambda_{ij}^{CA} + \lambda_{ik}^{CM} + \lambda_{jk}^{AM}$$

Postać multiplikatywna modelu:

$$\hat{n}_{ijk} = \eta \cdot \eta_i^C \cdot \eta_j^A \cdot \eta_k^M \cdot \eta_{ij}^{CA} \cdot \eta_{ik}^{CM} \cdot \eta_{jk}^{AM},$$

gdzie $\eta_i^X = exp(\lambda_i^X)$

Z elementów składowych modelu możemy wyczytać informacje o zależnościach między zmiennymi. Przejrzę teraz parametru modelu multiplikatywnego.

```
exp(mod2$param$cigarette)
```

```
##        no       yes
## 0.7540653 1.3261450
```

```
exp(mod2$param$alcohol)
```

```
##        no       yes
## 0.2222403 4.4996344
```

```
exp(mod2$param$marijuana)
```

```
##        no       yes
## 3.3070224 0.3023868
```

Interpretacja parametrów:

- **cigarette**
  - parametry odpowiadają efektom η_1^C, η_2^C
  - liczebność w komórce yes będzie o ok 33% większa od wartości bazowej
  - liczebność w komórce no będzie o ok 25% mniejsza od wartości bazowej
  - Wskaźnik 1,32 przy yes wskazuje na wpływ dodatni, co oznacza, że palenie papierosów ma wpływ stymulujący na zażywanie innych używek
  - Wskaźnik 0,75 przy no wskazuje na wpływ ujemny, czyli niepalenie papierosów ma ograniczajacy wpływ na stosowanie innych używek
- **alcohol**
  - parametry odpowiadają efektom η_1^A, η_2^A
  - liczebność w komórce yes będzie ok 4,5 razy większa od wartości bazowej
  - liczebność w komórce no będzie o ok 78% mniejsza od wartości bazowej
  - Wskaźnik 4,5 przy yes wskazuje na wpływ dodatni, czyli spożywanie alkoholu ma wpływ stymulujący na zażywanie innych używek. Wskaźnik jest znacznie wyższy od 1, co oznacza, że osoby spożywające alkohol dużo częściej sięgają także po inne używki
  - Wskaźnik 0,22 przy no wskazuje na wpływ ujemny, czyli spożywanie alkoholu ma ograniczajacy wpływ na stosowanie innych używek
- **marijuana**
  - parametry odpowiadają efektom η_1^M, η_2^M
  - liczebność w komórce yes będzie o ok 70% mniejsza od wartości bazowej
  - liczebność w komórce no będzie ok 3,3 razy większa od wartości bazowej
  - Wskaźnik 0,3 przy yes wskazuje na wpływ ujemny, co oznacza, że palenie marihuany ma ograniczający wpływ na stosowanie innych używek
  - Wskaźnik 3,3 przy no wskazuje na wpływ dodatni, czyli niepalenie marihuany ma wpływ stymulujący na zażywanie innych używek. Wskaźnik jest znacznie wyższy od 1, co oznacza, że osoba niepaląca będzie prawdopodobnie stosowała inne używki

```
sum(dane$cigarette=="no")
```

```
## [1] 781
```

```
sum(dane$alcohol=="no")
```

```
## [1] 327
```

```
sum(dane$marijuana=="no")
```

```
## [1] 1316
```

Część licealistów nie stosujących używek:

- 34% nie paliło papierosów
- 14% nie spożyło alkoholu
- 58% nie paliło marihuany

```
exp(mod2$param$cigarette.marijuana)
```

```
##          marijuana
## cigarette        no       yes
##       no  2.0380050 0.4906759
##       yes 0.4906759 2.0380050
```

```
exp(mod2$param$cigarette.alcohol)
```

```
##          alcohol
```

```
## cigarette        no       yes
##       no  1.6713421 0.5983216
##       yes 0.5983216 1.6713421
```

```
exp(mod2$param$marijuana.alcohol)
```

```
##          alcohol
## marijuana        no       yes
##       no  2.1096196 0.4740191
##       yes 0.4740191 2.1096196
```

Interpretacja parametrów:

- **cigarette & marijuana**
  - Wskaźnik 2,04 oznacza wpływ dodatni (yes,yes/no,no)
    - Jeśli dana osoba zażyła marihuanę i papierosy, to prawdopodobnie także spożywała alkohol
    - Jeśli dana osoba nie miała styczności z papierosami i marihuaną, to prawdopdopodobnie nie spozywała także alkoholu.
  - Wskaźnik 0,49 oznacza wpływ ujemny (yes,no)
    - Jeśli dana osoba stosuje tylko jedną z używek (papierosy, marihuana), to rzadziej będzie sięgała po alkohol
- **cigarette & alcohol**
  - Wskaźnik 1,67 oznacza wpływ dodatni (yes,yes/no,no)
    - Jeśli dana osoba zażyła alkohol i papierosy, to częściej sięgnie po marihuanę
    - Jeśli dana osoba nie miała styczności z papierosami i alkoholem, to prawdopdopodobnie jest zainteresowana także marihuaną.
  - Wskaźnik 0,6 oznacza wpływ ujemny (yes,no)
    - Jeśli dana osoba stosuje tylko jedną z używek (papierosy, alkohol), to rzadziej będzie sięgała po marihuanę
- **marijuana & alcohol**
  - Wskaźnik 2,11 oznacza wpływ dodatni (yes,yes/no,no)
    - Jeśli dana osoba zażyła marihuanę i alkohol, to prawdopodobnie także paliła papierosy
    - Jeśli dana osoba nie miała styczności z alkoholem i marihuaną, to prawdopdopodobnie nie paliła papierosów.
  - Wskaźnik 0,47 oznacza wpływ ujemny (yes,no)
    - Jeśli dana osoba stosuje tylko jedną z używek (alkohol, marihuana), to rzadziej będzie sięgała po papierosy

Tabela kontygencji:

- Osoby, które paliły papierosy i marihuanę:
  - 99,7% z tych osób spożywało także alkohol
- Osoby, które paliły papierosy i piły alkohol:
  - niecałe 63% z tych osób paliło także marihuanę
- Osoby, które paliły marihuanę i piły alkohol:
  - ponad 95% z tych osób paliło też papierosy

Zauważam, że najczęściej licealiści mieli styczność ze wszystkimi trzema używkami. Sporej większości zdarzyło się wypić alkohol, a jeśli nie spożywali alkoholu to zwykle nie sięgali także po inne używki. Może to wynikać z tego, że alkohol jest najbardziej dostępną z podanych używek. Ponadto jest najbardziej akceptowalny społecznie.