Modele log-liniowe chierachicznie uporządkowane

Aleksandra Palka

28 stycznia 2024

1 Wstęp

Projekt oparty jest na modelach log-liniowych chierarchicznie uporządkowanych. Będziemy pracować na danych zawierających odpowiedzi na pytania pewnej ankiety. Zawierała ona trzy pytania, które dotyczyły jakości snu (odpowiedź 1 oznaczała, że student sypia dobrze, 0, że źle), czy regularnie biega (1 – tak, 0 – nie) oraz czy posiada psa (1 – tak, 0– nie). Zmienną sen oznaczymy przez 1, Bieganie-2 oraz Pies-3.

2 Zadanie 1 - modele log-liniowe hierarchicznie uporządkowane

Podamy interpretacje następujących modeli log-liniowych hierarchicznie uporządkowych w oparciu o dane z *Ankieta.csv* i zapiszemy je w parametryzacji ANOVA.

2.1 model [1 3]

$$l_{ijk} = \lambda + \lambda_i^{(1)} + \lambda_k^{(3)}$$

Zmienne Sen(1) i Pies(3) są niezależne i mają dowolny rozkład. Zmienna Bieganie(2) ma równomierny rozkład i jest niezależna od zmiennych 1 i 3.

2.2 model [13]

$$l_{ijk} = \lambda + \lambda_i^{(1)} + \lambda_k^{(3)} + \lambda_{ik}^{(13)}$$

Zmienne Sen(1) i Pies(3) są nie są niezależne i mają dowolny rozkład. Zmienna Bieganie(2) ma równomierny rozkład i jest niezależna od zmiennych 1 i 3.

2.3 model [1 2 3]

$$l_{ijk} = \lambda + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)}$$

Zmienne Sen(1) i Bieganie(2) Pies(3) są między sobą niezależne.

2.4 model [12 3]

$$l_{ijk} = \lambda + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{ij}^{(12)}$$

Zmienne Sen(1) i Bieganie(2) są nie są niezależne. Zmienna Bieganie(2) jest niezależna od zmiennych 1 i 3.

2.5 model [12 13]

$$l_{ijk} = \lambda + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{ij}^{(12)} + \lambda_{ik}^{(13)}$$

Zmienne Pies(2) i Bieganie(2) są warunkowo niezależne(czyli są niezależne przy ustalonej zmiennej Sen(1)).

2.6 model [1 23]

$$l_{ijk} = \lambda + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{ij}^{(12)}$$

Zmienne Pies(3) i Bieganie(2) są nie są niezależne. Zmienna Sen(1) jest niezależna od zmiennych 2 i 3.

3 Zadanie 2 - szacowanie prawdopodobieństw zdarzeń

Oszacujemy pewne prawdopodobieństwa warunkowe na podstawie danych z *Ankieta.csv*. Do tego przyjmiemy model [12 3] i porównami z wynikami uzyskamy przy przyjęciu modelu [12 23]. Modele będziemy dopasowywać za pomocą funkcji *glm*.

```
ankieta <- read.csv2(file="Ankieta.csv",header=TRUE)
#model [12 3]
model1 <- glm(Freq ~ SEN + BIEGANIE + PIES +
                SEN*BIEGANIE,
              data = ankieta.df, family = poisson)
#model [12 23]
model2 <- glm(Freq ~ SEN + BIEGANIE + PIES +
                SEN*BIEGANIE+BIEGANIE*PIES,
              data = ankieta.df, family = poisson)
```

Porównanie liczności z danych i uzyskanych za pomocą modelu:

```
>cbind(model1$data, fitted(model1),fitted(model2))
  SEN BIEGANIE PIES Freq fitted(model1) fitted(model2)
1   0        0    0    6          3.400      4.8888889
2   1        0    0    5          4.250      6.1111111
3   0        1    0    1          1.275      0.8181818
4   1        1    0    5          8.075      5.1818182
5   0        0    1    2          4.600      3.1111111
6   1        0    1    5          5.750      3.8888889
7   0        1    1    2          1.725      2.1818182
8   1        1    1   14         10.925     13.8181818
>
```

3.1 Prawdopodobieństwo dobrej jakości snu studenta, który regularnie biega

Chcemy oszacować:

$$P(S = 1|B = 1) = \frac{P(S = 1 \wedge B = 1)}{P(B = 1)}$$

```
licznosciModel1 = cbind(model1$data, fitted(model1))[[5]]
licznosciModel2 = cbind(model2$data, fitted(model2))[[5]]
licznosci = model1$data[[4]]
pa_model1 = (licznosciModel1[4]+licznosciModel1[8])/(licznosciModel1[4]+
        licznosciModel1[8]+licznosciModel1[3]+licznosciModel1[7])
pa_model2 = (licznosciModel2[4]+licznosciModel2[8])/(licznosciModel2[4]+
        licznosciModel2[8]+licznosciModel2[3]+licznosciModel2[7])
pa_ = (licznosci[4]+licznosci[8])/(licznosci[4]+licznosci[8]+
        licznosci[3]+licznosci[7])
```

W tabeli 1 przedstawione są wartości oszacowanego prawdopodobieństwa. Dla każdego sprawdzanego modelu są one równe prawdopodobieństwu wyliczonemu z danych, ponieważ oba te modele uwzględniają zależności między sprawdzanymi zmiennymi Pies i Bieganie (występują w nich efekty $\lambda^{(12)}$).

	$P(S=1\|B=1)$
model [12 3]	0.8636364
model [12 23]	0.8636364
z danych	0.8636364

Tabela 1: Oszacowane prawdopodobieństwa dla 2.1

3.2 Prawdopodobieństwo tego, że student biega regularnie, gdy posiada psa.

Chcemy oszacować:

$$P(B=1|P=1) = \frac{P(B=1 \wedge P=1)}{P(P=1)}$$

```
pb_model1 = (licznosciModel1[7]+licznosciModel1[8])/(licznosciModel1[5]
        +licznosciModel1[8]+licznosciModel1[6]+licznosciModel1[7])
pb_model2 = (licznosciModel2[7]+licznosciModel2[8])/(licznosciModel2[5]+
        licznosciModel2[8]+licznosciModel2[6]+licznosciModel2[7])
pb_ = (licznosci[7]+licznosci[8])/(licznosci[6]+licznosci[8]+licznosci[5]+
        licznosci[7])
```

	$P(B=1\|P=1)$
model [12 3]	0.55
model [12 23]	0.6956522
z danych	0.6956522

Tabela 2: Oszacowane prawdopodobieństwa dla 2.2

Wyniki obliczeń widoczne są w tabeli 2. Badamy zależnośc zmiennych Bieganie i Pies i z modelu, który ją uwzględnia([12 23]) otrzymaliśmy takie samo prawdopodobieństwo jak z danych. Dla modelu [12 3], który zakłada niezależność zmiennej Pies od pozostałych oszacowany wynik wyraźnie się rózni.

4 Zadanie 3 - testowanie modeli

Będziemy weryfikować hipotezy testując określone modele log-liniowe przeciw modelowi pełnemu([123]) oraz innemy wybranemu nadmodelowi na poziomie istotności $\alpha = 0.05$. Statystyką testową jest G^2:

$$G^2 = 2\sum_{i=1}^{I} Y_i \ln\left(Y_i/n\hat{p_i}\right),$$

gdzie I-liczba komórek w tablicy liczności, Y_i-liczności w tej tabeli, a $\hat{p_i}$ to estymowane prawdopodobieństwa na podstawie modelu. Wartość G^2 uzyskujemy z funkcji *anova*. P-wartośc tego testu wyznaczamy ze wzoru:

$$p = 1 - F_{\chi^2_{I-q-1}}(G^2),$$

gdzie q-liczba estymowanych parametrów $\mathcal{M}^0$, $F_{\chi^2_{I-q-1}}$ jest dystrybuantą rozkładu χ^2 z $I-q-1$ stopniami swobody.

4.1 Zmienne losowe Sen, Bieganie i Pies są wzajemnie niezależne.

Niezależność wszystkich trzech zmiennych odpowiada modelowi [1 2 3].

$$H_0 : \mathcal{M}^0 = [1\ 2\ 3]$$

$$H_1 : \mathcal{M} = [123]$$

$$H_1 : \mathcal{M} = [12\ 23\ 31]$$

```
model_a <- glm(Freq ~ SEN + BIEGANIE + PIES ,
              data = ankieta.df, family = poisson)
model_h1 <- glm(Freq ~ (SEN + BIEGANIE + PIES)^2 ,
               data = ankieta.df, family = poisson)
# p-wartosc [123]
1-pchisq(deviance(model_a), df = df.residual(model_a))

test_a <- anova(model_a,model_h1)
#p-wartosc [12 23 31]
1-pchisq(test_a$Deviance[2], df = test_a$Df[2])
```

H_1	p-wartość
[123]	0.02932791
[12 23 31]	0.01438801

Tabela 3: P-wartości dla 3.1

Otrzymane p-wartości (tabela 3) dla obu testowanych hipotez alternatywnych są mniejsze od przyjętego poziomu istotności. Zatem odrzucamy hipotezę zerową-model [1 2 3] nie jest dobrym dopasowaniem- zmienne Sen, Bieganie i Pies nie są wzajemnie niezależne. Modele [123] i [12 23 31] lepiej opisują dane.

4.2 Zmienna losowa Pies jest niezależna od pary zmiennych Sen i Bieganie.

Będziemy testować model [12 3].

$$H_0 : \mathcal{M}^0 = [12\ 3]$$

$$H_1 : \mathcal{M} = [123]$$

$$H_1 : \mathcal{M} = [12\ 23\ 31]$$

```
model_b <- glm(Freq ~ SEN + BIEGANIE + PIES + SEN*BIEGANIE,
               data = ankieta.df, family = poisson)
#p-wartosc [123]
1-pchisq(deviance(model_b), df = df.residual(model_b))

test_b <- anova(model_b,model_h1)
#p-wartosc [12 23 31]
1-pchisq(test_b$Deviance[2], df = test_b$Df[2])
```

H_1	p-wartość
[123]	0.1131637
[12 23 31]	0.05618272

Tabela 4: P-wartości dla 3.2

Dla testów przeciwko obu modelom w hipotezach alternatywnych p-wartość (tabela 4) jest większa od 0.05. Na tym poziomie istotności nie mamy podstaw do odrzucenia hipotezy zerowej o niezależności zmiennej Pies od zmiennych Sen i Bieganie.

4.3 Zmienna losowa Sen jest niezależna od zmiennej Pies, przy ustalonej zmiennej Bieganie.

Przetestujemy model [12 23].

$$H_0 : \mathcal{M}' = [12\ 23]$$

$$H_1 : \mathcal{M} = [123] \ i \ \mathcal{M} \neq [12\ 23]$$
$$H_1 : \mathcal{M} = [12\ 23\ 31] \ i \ \mathcal{M} \neq [12\ 23]$$

Przy takiej postaci hipotezy alternatywnej:

$$G^(\mathcal{M}^0|\mathcal{M}) = (\mathcal{M}^0|\mathcal{M}_d) - (\mathcal{M}^0|\mathcal{M}_d),$$

gdzie $\mathcal{M}_d$ jest modelem pełnym. Wtedy

$$p = 1 - F_{\chi^2_{r-s}}(G^2(\mathcal{M}^0|\mathcal{M})),$$

gdzie r-liczba stopni swobody w teście $\mathcal{M}^0$ przeciwko $\mathcal{M}_d$, a s-liczba stopni swobody w teście $\mathcal{M}$ przeciwko $\mathcal{M}_d$.

```
model_c <- glm(Freq ~ SEN*BIEGANIE + PIES*BIEGANIE,
               data = ankieta.df, family = poisson)
model_pelny <- glm(Freq ~ (SEN + BIEGANIE + PIES)^2 +SEN*BIEGANIE*PIES,
                   data = ankieta.df, family = poisson)
#p-wartosc model [123]
1-pchisq(deviance(model_c)-deviance(model_pelny), df = df.residual(model_c)-df.residual(model_pelny))
#p-wartosc model [12 23 31]
1-pchisq(deviance(model_c)-deviance(model_h1), df = df.residual(model_c)-df.residual(model_h1))
```

H_1	p-wartość
[123]	0.5329187
[12 23 31]	0.3057874

Tabela 5: P-wartości dla 3.3

W tabeli 5 p-wartości są większe od zadanego poziomu istotności. Nie mamy podstaw do odrzucenia hipotezy zerowej, przy ustalonej zmiennej Bieganie, zmienna Sen jest niezależna od zmiennej Pies.