Politechnika Wrocławska
Wydział Matematyki

Skład grupy:	Agata Sobczak 268873
	Jakub Franczak 262271
	Katarzyna Kudełko 268762
Prowadząca laboratorium:	dr inż. Aleksandra Grzesiek
Prowadząca wykładu:	dr hab. Alicja Jokiel-Rokita

Analiza Danych Ankietowych

Raport 4.

Lista 4.

WROCŁAW 29.1.2024

Spis treści

1 Zadanie 1

1.1 Cel zadania

Celem zadania jest podanie interpretacji podanych modeli log - liniowych.

(a) [1 3]

Model ten zakłada niezależność zmiennych 1 i 3. Można go zapisać:

$$\ell_{ik} = \lambda + \lambda_i^{(1)} + \lambda_k^{(3)}, \tag{1}$$

gdzie $\forall i \in \{1, \dots, R\}$, $k \in \{1, \dots, L\}$.

(b) [13]

Model ten zakłada zależność zmiennych 1 i 3. Można go zapisać:

$$\ell_{ik} = \lambda + \lambda_i^{(1)} + \lambda_k^{(3)} + \lambda_{ik}^{(13)}, \tag{2}$$

gdzie $\forall i \in \{1, \dots, R\}$, $k \in \{1, \dots, L\}$.

(c) [1 2 3]

Model ten oparty jest na trzech zmiennych i zakłada ich niezależność względem siebie. Można go zapisać:

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

gdzie $\forall i \in \{1, \dots, R\}$, $j \in \{1, \dots, C\}$, $k \in \{1, \dots, L\}$.

(d) [12 3]

Model ten oparty jest na trzech zmiennych i zakłada, że zmienna 1 i 2 są od siebie zależne, jednak brakuje brak zależności wobec zmiennej 3. Można go zapisać:

$$\ell_{ijk} = \lambda + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{ij}^{(12)}, \tag{4}$$

gdzie $\forall i \in \{1, \dots, R\}$, $j \in \{1, \dots, C\}$, $k \in \{1, \dots, L\}$.

(e) [12 13]

Model ten oparty jest na trzech zmiennych i zakłada, że zmienne 1 i 2, a także 1 i 3 są od siebie zależne. Przy ustalonej wartości zmiennej 1, zmienne 2 i 3 stają się wzajemnie niezależne. W takim przypadku można opisać zmienne 2 i 3 jako niezależne warunkowo.

$$\ell_{ijk} = \lambda + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{ij}^{(12)} + \lambda_{ik}^{(13)}, \tag{5}$$

gdzie $\forall i \in \{1, \dots, R\}$, $j \in \{1, \dots, C\}$, $k \in \{1, \dots, L\}$.

(f) [1 23]

Model ten oparty jest na trzech zmiennych i zakłada, że zmienna 1 jest niezależna od zmiennych 2 i 3, ale zmienne 2 i 3 są od siebie zależne. Można go zapisać:

$$\ell_{ijk} = \lambda + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{jk}^{(23)}, \tag{6}$$

gdzie $\forall i \in \{1, \dots, R\}$, $j \in \{1, \dots, C\}$, $k \in \{1, \dots, L\}$.

2 Zadanie 2

2.1 Cel zadania

Celem zadania jest oszacowanie prawdopodobieństw podanych poniżej, na podstawie danych *Ankieta.csv* przyjmując model log- normalny [12 3] oraz sprawdzenie jakie byłyby oszacowania tych prawdopodobieństw przy założeniu modelu [12 23].

- a) dobrej jakosci snu studenta, który regularnie biega,
- b) tego, ze student biega regularnie, gdy posiada psa.

2.2 Modele

W tym zadaniu zakładamy, że

- zmienna 1 to zmienna S - jakość snu studenta (0 - zła jakość, 1 - dobra jakość)
- zmienna 2 to zmienna B - regularne bieganie (0 - nie biega regularnie, 1 - biega regularnie)
- zmienna 3 to zmienna P - posiadanie psa (0 - nie posiada psa, 1 - posiada psa)

Pierwszy model oparty jest na trzech wymienionych zmiennych i zakłada, że zmienna 1 i 2 są od siebie zależne i nie ma zależności wobec zmiennej 3.

$$\ell_{ijk} = \lambda + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{ij}^{(12)}, \tag{7}$$

gdzie $\forall i \in \{0,1\}$, $j \in \{0,1\}$, $k \in \{0,1\}$.
Oznacza to, że zakładamy, że jakość snu jest powiązana z regularnym bieganiem, jednak posiadanie psa jest czynnikiem od nich niezależnym.

Drugi model oparty jest na trzech zmiennych i zakłada, że zmienne 1 i 2, a także 2 i 3 są od siebie zależne. Przy ustalonej wartości zmiennej 2, zmienne 1 i 3 stają się wzajemnie niezależne. W takim przypadku można opisać zmienne 1 i 3 jako niezależne warunkowo.

$$\ell_{ijk} = \lambda + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{ij}^{(12)} + \lambda_{jk}^{(23)}, \tag{8}$$

gdzie $\forall i \in \{0,1\}$, $j \in \{0,1\}$, $k \in \{0,1\}$.
Oznacza to, że zakładamy, że jakość snu i regularne bieganie są od siebie zależne jak i regularnie bieganie i posiadania psa.
Prawdopodobieństwa szacujemy na podstawie poniższych modeli:

```
mod1 <- glm(Freq ~ sen + bieganie + pies + sen * bieganie,
            data = ankieta.df, family = poisson)
```

```
mod2 <- glm(Freq ~ sen + bieganie + pies + sen * bieganie + bieganie * pies,
            data = ankieta.df, family = poisson)
```

2.3 a)

Model:	[12 3]	[12 13]
Prawdopodobieństwo	0.8636364	0.8636364

Prawdopodobieństwo dobrej jakosci snu studenta, który regularnie biega wynosi w przybliżeniu 86 % niezależnie od przyjętego modelu log-liniowego.

2.4 b)

Model:	[12 3]	[12 13]
Prawdopodobieństwo	0.6956522	0.6956522

Prawdopodobieństwo tego, ze student biega regularnie, gdy posiada psa wynosi w przybliżeniu 70 % niezależnie od przyjętego modelu log- liniowego.

3 Zadanie 3.

3.1 Cel zadania

Celem zadania jest zweryfikowanie następujących hipotez dotyczących parametrów modelu log-liniowego hierarchicznie uporządkowanego na podstawie danych z pliku Ankieta.csv, na poziomie istotności $\alpha = 0.05$.

- a) zmienne losowe Sen, Bieganie i Pies są wzajemnie niezależne,
- b) zmienna losowa Pies jest niezależna od pary zmiennych Sen i Bieganie,
- c) zmienna losowa Sen jest niezależna od zmiennej Pies, przy ustalonej zmiennej Bieganie.

3.2 a)

Ustalono H_0 – dane pochodzą z modelu [1 2 3].

```
# a)

mod1 <- glm(Freq ~ Sen + Bieganie + Pies,
            data = df_tab, family = poisson)

p1 <- 1-pchisq(deviance(mod1), df = df.residual(mod1))
p1
```

Rysunek 1: Kod badający model [1 2 3]

Otrzymana p-wartość wynosi 0.02932791, zatem jest mniejsza od danego poziomu istotności. Należy odrzucić hipotezę zerową o pochodzeniu danych z modelu [1 2 3], gdzie wszystkie badane zmienne są niezależne, na rzecz hipotezy alternatywnej.

Następnie ustalono H_0 – dane pochodzą z modelu [1 2 3], a H_1 - dane pochodzą z nadmodelu [13 32].

```
# nadmodel

mod2 <- glm( Freq ~ Sen + Bieganie + Pies + Sen*Pies + Pies*Bieganie, data =df_tab, family = poisson)
test <- anova(mod1, mod2)
p2 <- 1-pchisq(test$Deviance[2], df = test$Df[2])
p2
```

Rysunek 2: Kod badający model [1 2 3] przeciwko modelowi [13 32]

Ponownie uzyskana p-wartość jest mniejsza od poziomu istotności α, ponieważ wynosi ona 0.02381948. Istnieją zatem podstawy by odrzucić hipotezę zerową.

3.3 b)

Ustalono H_0 – dane pochodzą z modelu [12 3].

```
# b)

mod3 <- glm( Freq ~ Sen + Pies + Bieganie + Pies + Sen*Bieganie, data =df_tab, family = poisson)
p3 <- 1-pchisq(deviance(mod3), df = df.residual(mod3))
p3
```

Rysunek 3: Kod badający hipotezę H_0

Uzyskana p-wartość wynosi $0.1131637 > \alpha$, zatem nie ma podstaw do odrzucenia hipotezy o niezależności zmiennej losowej Pies od pary zmiennych Sen i Bieganie.

Następnie zbadano model [12 3] przeciwko nadmodelowi [13 32].

```
# nadmodel

mod4 <- glm( Freq ~ Sen + Pies + Bieganie + Sen*Pies + Bieganie*Pies, data=df_tab, family = poisson)
test <- anova(mod3,mod4)
p4 <- 1-pchisq(test$Deviance[2], df = test$Df[2])
p4
```

Rysunek 4: Kod badający model [12 3] przeciwko nadmodelowi [13 23]

Uzyskana p-wartość jest równa 0.101799. Jest ona większa niż zadany poziom istotności α, zatem nie ma podstaw do odrzucenia hipotezy zerowej.

3.4 c)

Ustalono H_0 – dane pochodzą z modelu [12 23].

```
# c)


mod5 <- glm( Freq ~ Sen + Pies + Bieganie + Sen*Bieganie + Pies*Bieganie, data =df_tab, family = poisson)
p5 <- 1-pchisq(deviance(mod5), df = df.residual(mod5))
p5
```

Rysunek 5: Kod badający hipotezę H_0

P-wartość dla tego testu wynosi 0.5329187. Jest to o wiele wyższa wartość niż przyjęty poziom istotności, dlatego można wnioskować przyjęcie hipotezy zerowej.

```
mod6 <- glm( Freq ~ Sen + Pies + Bieganie + Sen*Bieganie + Sen*Pies + Pies*Bieganie, data=df_tab, family = poisson)
test <- anova(mod5, mod6)
p6 <- 1-pchisq(test$Deviance[2], df = test$Df[2])
p6
```

Rysunek 6: Kod badający model [12 23] przeciwko nadmodelowi [12 13 23]

W badaniu modelu [12 23] przeciwko nadmodelowi [12 13 23] otrzymana p-wartość to 0.3057874. Ponownie nie ma podstaw do odrzucenia hipotezy.

4 Zadanie 4.

4.1 Cel zadania

Celem zadania jest dokonanie wyboru modelu na podstawie danych Ankieta.csv w oparciu o kryterium AIC i kryterium BIC.

4.2 Kody

```
ankieta <- array(data = c(6, 5, 1, 5, 2, 5, 2, 14),
                 dim = c(2,2,2),
                 dimnames = list("sen" = c("0","1"),
                                 "pies" = c("0","1"),
                                 "bieganie" = c("0","1")))

addmargins(ankieta)
ankieta.df <- as.data.frame(as.table(ankieta))
ankieta.df[,-4] <- lapply(ankieta.df[,-4], relevel, ref = "0")
ankieta.df
```

Rysunek 7: Inicjalizacja tabeli danych

```
mod1 <- glm(Freq ~ 1,
            data = ankieta.df, family = poisson)
mod2 <- glm(Freq ~ sen,
            data = ankieta.df, family = poisson)
mod3 <- glm(Freq ~ pies,
            data = ankieta.df, family = poisson)
mod4 <- glm(Freq ~ bieganie,
            data = ankieta.df, family = poisson)
mod5 <- glm(Freq ~ sen + bieganie,
            data = ankieta.df, family = poisson)
mod6 <- glm(Freq ~ sen + pies,
            data = ankieta.df, family = poisson)
mod7 <- glm(Freq ~ pies + bieganie,
            data = ankieta.df, family = poisson)
mod8 <- glm(Freq ~ sen + pies + sen * pies,
            data = ankieta.df, family = poisson)
mod9 <- glm(Freq ~ sen + bieganie + sen * bieganie,
            data = ankieta.df, family = poisson)
mod10 <- glm(Freq ~ pies + bieganie + pies * bieganie,
             data = ankieta.df, family = poisson)
mod11 <- glm(Freq ~ sen + pies + bieganie + pies * bieganie,
             data = ankieta.df, family = poisson)
mod12 <- glm(Freq ~ sen + bieganie + pies + sen*bieganie,
             data = ankieta.df, family = poisson)
mod13 <- glm(Freq ~ sen + pies + bieganie + sen*pies,
             data = ankieta.df, family = poisson)
mod14 <- glm(Freq ~ sen + pies + bieganie + sen*bieganie + bieganie*pies,
             data = ankieta.df, family = poisson)
mod15 <- glm(Freq ~ sen + pies + bieganie +  sen*pies + sen*bieganie,
             data = ankieta.df, family = poisson)
mod16 <- glm(Freq ~ sen + pies + bieganie + sen*pies + pies * bieganie,
             data = ankieta.df, family = poisson)
mod17 <- glm(Freq ~ sen + pies + bieganie,
             data = ankieta.df, family = poisson)
mod18 <- glm(Freq ~ sen + pies + bieganie + sen*pies + sen*bieganie + pies*bieganie,
             data = ankieta.df, family = poisson)
mod19 <- glm(Freq ~ sen + pies + bieganie + sen*pies + sen*bieganie + pies*bieganie + sen * pies * bieganie,
             data = ankieta.df, family = poisson)
```

Rysunek 8: Badanie modeli

4.3 Wyniki i wnioski

```
> print(wyniki)
    Model      AIC      BIC
1   Mod 1 48.28345 48.36289
2   Mod 2 41.88518 42.04406
3   Mod 3 49.88278 50.04166
4   Mod 4 49.38005 49.53893
5   Mod 5 42.98177 43.22010
6   Mod 6 43.48451 43.72283
7   Mod 7 50.97938 51.21770
8   Mod 8 40.68714 41.00490
9   Mod 9 42.21678 42.53454
10 Mod 10 48.26987 48.58763
11 Mod 11 41.87160 42.26880
12 Mod 12 43.81611 44.21332
13 Mod 13 41.78373 42.18094
14 Mod 14 41.10660 41.58325
15 Mod 15 41.01874 41.49539
16 Mod 16 39.07422 39.55087
17 Mod 17 44.58110 44.89887
18 Mod 18 40.02544 40.58153
19 Mod 19 41.81545 42.45098
```

```
Start:  AIC=41.82
Freq ~ sen + pies + bieganie + sen * pies + sen * bieganie +
    pies * bieganie + sen * pies * bieganie

                    Df Deviance    AIC
- sen:pies:bieganie  1  0.20999 40.025
<none>                  0.00000 41.815

Step:  AIC=40.03
Freq ~ sen + pies + bieganie + sen:pies + sen:bieganie + pies:bieganie

                Df Deviance    AIC
- sen:bieganie   1   1.2588 39.074
<none>               0.2100 40.025
- pies:bieganie  1   3.2033 41.019
- sen:pies       1   3.2912 41.107

Step:  AIC=39.07
Freq ~ sen + pies + bieganie + sen:pies + pies:bieganie

                Df Deviance    AIC
<none>               1.2588 39.074
- pies:bieganie  1   5.9683 41.784
- sen:pies       1   6.0561 41.872
```

```
Call:  glm(formula = Freq ~ sen + pies + bieganie + sen:pies + pies:bieganie,
    family = poisson, data = ankieta.df)

Coefficients:
    (Intercept)           sen1          pies1      bieganie1     sen1:pies1  pies1:bieganie1
         1.5870         0.2231        -1.7876        -0.4520         1.6227           1.4328

Degrees of Freedom: 7 Total (i.e. Null);  2 Residual
Null Deviance:      20.47
Residual Deviance: 1.259        AIC: 39.07
```

Rysunek 9: Otrzymane wyniki

Kryteria AIC (Akaike Information Criterion) i BIC (Bayesian Information Criterion) to metody oceny modeli statystycznych, które szacują równowagę między dokładnością modelu a jego złożonością. Zasadniczo, niższe wartości tych kryteriów wskazują na lepsze modele. Gdy analizuje się różne modele z uwzględnieniem różnorodnych kombinacji zmiennych, można zauważyć wahania w wartościach AIC i BIC. Model 16 odznacza się najniższymi wartościami w obu tych kryteriach. Sugeruje to, że ten model zapewnia najbardziej optymalne dopasowanie.

Dodatkowo, używając metody selekcji krokowej, stwierdzamy, że model o wartości AIC równej 39.07 i Residual Deviance (miara dopasowania modelu do danych) wynoszącej 1.259, również okazuje się być najbardziej optymalny. Jest to wspomniany wcześniej model 16, co potwierdza jego skuteczność w zakresie balansowania między złożonością a zdolnością do adekwatnego przedstawienia danych.