

מודלים סטטיסטיים ויישומיהם 52518 תשע"ח – פתרון תרגיל 13

1. נניח כי נתונם לנו 3 מודלים לוג-לינאריים המקבנים זה בזה, $M_1 \subset M_2 \subset M_3$. נסמן באופן $(M_r|M_s)$ את הסטטיסטי G^2 לבחינת המודל M_r מול המודל M_s , כאשר $r > s$. הוכחו כי מתקיים:
- $$G^2(M_1|M_3) = G^2(M_1|M_2) + G^2(M_2|M_3)$$
- (תבונה זו מוכרת בתור האדיטיביות של הסטטיסטי G^2).

$$\begin{aligned} G^2(M_a|M_b) &= 2 \sum_{i,j,k} F_{ijk} \log \left(\frac{\hat{F}_{ijk}^{(b)}}{\hat{F}_{ijk}^{(a)}} \right), \quad M_a \subset M_b \\ G^2(M_1|M_2) + G^2(M_2|M_3) &= 2 \sum_{i,j,k} F_{ijk} \log \left(\frac{\hat{F}_{ijk}^{(2)}}{\hat{F}_{ijk}^{(1)}} \right) + 2 \sum_{i,j,k} F_{ijk} \log \left(\frac{\hat{F}_{ijk}^{(3)}}{\hat{F}_{ijk}^{(2)}} \right) \\ &= 2 \sum_{i,j,k} F_{ijk} \left(\log \left(\frac{\hat{F}_{ijk}^{(2)}}{\hat{F}_{ijk}^{(1)}} \right) + \log \left(\frac{\hat{F}_{ijk}^{(3)}}{\hat{F}_{ijk}^{(2)}} \right) \right) = 2 \sum_{i,j,k} F_{ijk} \left(\log \left(\frac{\hat{F}_{ijk}^{(2)}}{\hat{F}_{ijk}^{(1)}} \cdot \frac{\hat{F}_{ijk}^{(3)}}{\hat{F}_{ijk}^{(2)}} \right) \right) \\ &= 2 \sum_{i,j,k} F_{ijk} \left(\log \left(\frac{\hat{F}_{ijk}^{(3)}}{\hat{F}_{ijk}^{(1)}} \right) \right) = G^2(M_1|M_3) \end{aligned}$$

2. נחזר לננתוני **self-esteem** שהוצגו בכיתה (מצורפים) ונעסק בעת בננתוני הבנים בלבד. בחנו את המודל עם פרמטרי ג'סן ראשוני עבור המשתנים GPA, גזע והערכה עצמית (וללא פרמטרי ג'סן נוספים). עבור מודל זה, חשבו דגנית את $\hat{\pi}_{ijk}$ ואת הסטטיסטי G^2 לבחינת המודל הנ"ל מול המודל המלא. השוו את תוצאותיכם מול התוצאות המתוקלות מהרצת הפונקציה `logm`.

הנתונים ($N = 142$):

GPA	Race	Esteem	Count	\hat{p}_{ijk}
1	1	1	15	0.1056
1	1	2	9	0.0633
1	2	1	17	0.1197
1	2	2	10	0.0704
2	1	1	26	0.1831
2	1	2	17	0.1197
2	2	1	22	0.1549
2	2	2	26	0.1831

$$\hat{\pi}_{1..} = 0.3592, \quad \hat{\pi}_{.1.} = 0.4718, \quad \hat{\pi}_{..1} = 0.5634$$

$$\text{בשים לב כי } \hat{F}_{ijk}^{(2)} / \hat{F}_{ijk}^{(1)} = \hat{p}_{ijk} / \hat{\pi}_{ijk}$$

GPA	Race	Esteem	Count	$\hat{\pi}_{ijk}$	$\hat{p}_{ijk} / \hat{\pi}_{ijk}$	$\log(\hat{p}_{ijk} / \hat{\pi}_{ijk})$	$F_{ijk} \log(\hat{p}_{ijk} / \hat{\pi}_{ijk})$
1	1	1	15	0.0955	1.1061	0.1009	1.5128
1	1	2	9	0.074	0.8565	-0.1549	-1.3942
1	2	1	17	0.1069	1.1199	0.1132	1.9252
1	2	2	10	0.0828	0.8505	-0.1619	-1.6191
2	1	1	26	0.1703	1.0752	0.0725	1.884
2	1	2	17	0.132	0.907	-0.0977	-1.6602
2	2	1	22	0.1907	0.8124	-0.2077	-4.5701
2	2	2	26	0.1478	1.2388	0.2142	5.5683

ובอก הכל קיבל: $G^2 = 3.2958$, בהערכתה של $\log_{10} \text{מתקבל}$ מתקבל $G^2 = 2 \cdot 1.6467 = 3.2933$

```
> library(MASS)
> D <- read.table("self-esteem-dat.txt", header = T, nrow = 8)
> loglm(count ~ GPA + Race + Esteem, data = D)
Call:
loglm(formula = count ~ GPA + Race + Esteem, data = D)
```

statistics:

	X ²	df	P(> X ²)
Likelihood Ratio	3.295836	4	0.5095922
Pearson	3.301507	4	0.5086935

3. בשנות ה-70 נערכ בארה"ב סקר עם השאלה "האם אתה מסכימ מהמשפט 'רצו שנשים ינהלו את משק הבית וישאירו את ניהול המדינה לגברים?'". נתוני הסקר מופיעים לפניכם (ובקובץ targ13dat.txt). עליכם למצוא את המודל הטוב ביותר ולתת לו אינטראקטיבית מתאימה.

Subjects in the 1974 General Social Survey, Cross-Classified by Attitude Toward Women, Staying at Home, Sex of Respondent, and Education of Respondent^a

Sex of respondent	Education of respondent in years ^b	Response				Total number	
		Agree		Disagree			
		No.	Percent.	No.	Percent.		
Male	≤ 8	89	67.4	43	32.6	132	
	9–12	102	35.9	182	64.1	284	
	≥ 13	48	19.9	193	80.1	241	
	Total	239	36.4	418	63.6	657	
Female	≤ 8	83	74.1	29	25.9	112	
	9–12	152	34.9	284	65.1	436	
	≥ 13	33	14.8	190	85.2	223	
	Total	268	34.8	503	65.2	771	
Total	≤ 8	172	70.5	72	29.5	244	
	9–12	254	35.3	466	64.7	720	
	≥ 13	81	17.5	383	82.5	464	
	Total	507	35.7	921	64.3	464	

Subjects in the 1975 General Social Survey, Cross-Classified by Attitude toward Women Staying at Home, Sex of Respondent, and Education of Respondent^a

Sex of respondent	Education of respondent, yrs.	Response ^b				Total no.	
		Agree		Disagree			
		No.	Percent.	No.	Percent.		
Male	≤ 8	72	60.5	47	39.5	119	
	9–12	110	35.9	196	64.1	306	
	≥ 13	44	19.7	179	80.3	223	
	Total	226	34.9	422	65.1	648	
Female	≤ 8	86	69.4	38	30.6	124	
	9–12	173	37.9	283	62.1	456	
	≥ 13	28	13.0	187	87.0	215	
	Total	287	36.1	508	63.9	795	
Total	≤ 8	158	65.0	85	35.0	243	
	9–12	283	37.1	479	62.9	762	
	≥ 13	72	16.4	366	83.6	438	
	Total	513	35.6	930	64.4	1443	

נסמן את המשתנים באופן הבא: A:year, B:sex, C:educ, D:opinion (עם כל האינטראקציות בבל הרמות):

```
> library(MASS)
> dat <- read.table("targ13dat.txt", header = T)
> df <- dat
> df$year <- factor(df$year)
```

(Dispersion parameter for poisson family taken to be 1)

```
Null deviance: 1.2157e+03  on 23  degrees of freedom
Residual deviance: 1.1058e-13  on  0  degrees of freedom
AIC: 201.11
```

Number of Fisher Scoring iterations: 3

ניתן לראות כי

- תרומתם של משתני האינטראקציה המרובעת אינן מובהקת, לבן נפסול את המודל ($ABCD$)
 - האינטראקציות המשולשות אינן מובהקות, למעט האינטראקציה BCD – מודל אפשרי יהיה (A, BCD)
 - האינטראקציות הזוגות המובהקות הן חסוכות educ,opinop1 (בלומר האינטראקציה CD) sex,educ-1 (בלומר האינטראקציה BC) – מודל אפשרי יהיה (A, BC, CD)

לפיכך, נבחן את המודלים $M_1: (A, BC, CD)$, $M_2: (A, BCD)$ ו- $M_3: (ABCD)$

עבור M_2

```
> loglm(freq ~ year + sex * educ * opinion, data = df)
```

call:

```
loglm(formula = freq ~ year + sex * educ * opinion, data = df)
```

Statistics:

x^2 df P(> x^2)

Likelihood Ratio 6.479094 11 0.8395647
Pearson 6.470421 11 0.8402077

עבור M_1

```
> loglm(freq ~ year + sex * educ + educ * opinion, data = df)
Call:
loglm(formula = freq ~ year + sex * educ + educ * opinion, data = df)
```

statistics:

	X ²	df	P(> X ²)
Likelihood Ratio	15.23999	14	0.3619486
Pearson	14.90021	14	0.3850099

בכומר, המודל (A, BC, CD) : M_1 נותן הסבר טוב מספיק לנחותים (לא דוחים את השערת האפס לפי ה-p-value שתקבל).

בעת נבדוק את המודלים M_4 : (A, B, CD) , M_5 : (A, BC, D) , M_6 : (A, B, C, D) –

עבור M_4

```
> loglm(freq ~ year + sex + educ * opinion, data = df)
Call:
loglm(formula = freq ~ year + sex + educ * opinion, data = df)
```

statistics:

	X ²	df	P(> X ²)
Likelihood Ratio	54.66588	16	4.034693e-06
Pearson	54.59050	16	4.151147e-06

עבור M_5

```
> loglm(freq ~ year + sex * educ + opinion, data = df)
Call:
loglm(formula = freq ~ year + sex * educ + opinion, data = df)
```

statistics:

	X ²	df	P(> X ²)
Likelihood Ratio	376.9377	16	0
Pearson	370.6325	16	0

עבור M_6

```
> loglm(freq ~ year + sex + educ + opinion, data = df)
Call:
loglm(formula = freq ~ year + sex + educ + opinion, data = df)
```

statistics:

	X ²	df	P(> X ²)
Likelihood Ratio	416.3636	18	0
Pearson	404.4172	18	0

ובסה"כ המודל הטוב ביותר הוא (A, BC, CD) : הדעה אינה תלולה במין המשיב ובשנת ערבית הסקר, אולם תלולה בהשכלה המשיב. כמו כן קיימת תלות בין מין ורמת השכלה.