

MATH 523 - EXERCISES 2

These exercises are not for assessment

- 1 For the Classroom Behaviour data from lectures, fit the 18 models that involve all possible combinations of main effects, two-way and three-way interactions for the three factor predictors. Carry out a complete analysis of deviance to identify the most appropriate model for the data.

Assume a log-linear model for the data, assumed to be independently Poisson distributed.

- 2 The following data are counts of observed hair and eye color combinations for a sample of female and male subjects.

		Female				Male			
Eyes		Blue	Brown	Green	Hazel	Blue	Brown	Green	Hazel
Hair	Black	9	36	2	5	11	32	3	10
	Blond	64	4	8	5	30	3	8	5
	Brown	34	81	14	29	50	38	15	25
	Red	7	16	7	7	10	10	7	7

The following code in R can be used to define the data in a data frame:

```
1 hair <- factor(rep(rep(c("Black", "Brown", "Red", "Blond"), c(4,4,4,4)),2))
2 eye <- factor(rep(c("Brown", "Blue", "Hazel", "Green"), 8))
3 sex <- factor(rep(c("Male", "Female"), c(16,16)))
4 freq <- c(32,11,10,3, 38,50,25,15, 10,10,7,7, 3,30,5,8,
5 36,9,5,2, 81,34,29,14, 16,7,7,7, 4,64,5,8)
6
7 haireye.data <- data.frame(hair, eye, sex, freq)
```

Carry out an analysis of deviance for this three way ($2 \times 4 \times 4$) table.

- 3 The following data represents counts of migration between four regions (NE-northeast, MW-midwest, S-south, W-west) of the USA between 1980 and 1985.

		1985			
		NE	MW	S	W
1980	NE	11607	100	366	124
	MW	87	13677	515	302
	S	172	225	17819	270
	W	63	176	286	10192

The following code in R can be used to define the data in a data frame:

```
1 count <- c(11607, 100, 366, 124, 87, 13677, 515, 302, 172, 225,
2 17819, 270, 63, 176, 286, 10192)
3 region <- c("NE", "MW", "S", "W")
4 row <- gl(4, 4, labels = region)
5 col <- gl(4, 1, length = 16, labels = region)
```

For this square, 4×4 table, a number of special log-linear models are of interest: for $j, k = 1, \dots, 4$,

M1 Independence: `row+col`

$$\log \mu_{jk} = \beta + \beta_j^{(R)} + \beta_k^{(C)}$$

M2 Quasi-Independence: `row+col+Diag(row, col)`,

$$\log \mu_{jk} = \beta + \beta_j^{(R)} + \beta_k^{(C)} + \gamma_j \mathbb{1}\{j = k\}$$

M3 Symmetry: `Symm(row, col)`

$$\log \mu_{jk} = \beta + \beta_j^{(R)} + \beta_k^{(C)} + \gamma_{jk}$$

where

- $\gamma_{jk} = \gamma_{kj}$ for $j \neq k$, and
- $\gamma_{jk} = 0, \beta_j^{(R)} = \beta_k^{(C)}$ for $j = k$.

M4 Quasi-Symmetry: `row+col+Symm(row, col)`

$$\log \mu_{jk} = \beta + \beta_j^{(R)} + \beta_k^{(C)} + \gamma_{jk}$$

where

- $\gamma_{jk} = \gamma_{kj}$ for $j \neq k$, and

Note that in these representations of the models, as for standard factor predictor models, some parameters may be aliased, necessitating the use of identifiability constraints.

The following code can be used to fit these models using the `gnm` library in R:

```
1 library(gnm)
2
3 independence <- glm(count ~ row + col, family = poisson)
4 quasi.indep <- glm(count ~ row + col + Diag(row, col), family = poisson)
5 symmetry <- glm(count ~ Symm(row, col), family = poisson)
6 quasi.symm <- glm(count ~ row + col + Symm(row, col), family = poisson)
7 comparison1 <- anova(independence, quasi.indep, quasi.symm)
8 print(comparison1)
```

See cran.r-project.org/web/packages/gnm/vignettes/gnmOverview.pdf

- (a) Assess which of the models is most appropriate for these data using an analysis of deviance and chi-squared tests. Note the numbers of parameters fitted in each model.
- (b) Use the R function `fitted` to inspect the fitted values for each model.
- (c) Explain the differences between the Symmetry and Quasi-Symmetry models in terms of the marginal implications of the models (that is, the implications for the distributions of the row and column totals).