MSc: Statistical Analysis (Spring Term)

Generalized Linear Models: Exercises 3 (log-linear models)

In these examples you will explore the use of R to model two-way (and higher) contingency tables, using a *log-linear* model. That is, a generalized linear model, assuming a *Poisson* distributed response variable, being the cell count of the table, and a *log* link.

Exercises

1. The table below shows the results of a skin cancer survey. For a sample of 400 patients, the SITE of the tumour, and its histological TYPE were recorded.

	SITE			
TUMOUR TYPE	Head and neck	Trunk	Extremities	**Total**
Hutchinson's melantonic freckle	22	2	10	34
Superficial spreading melanoma	16	54	115	185
Nodular	19	33	73	125
Intermediate	11	17	28	56
Total	68	106	226	400

   (a) Assuming that both SITE and TYPE are response factors (i.e. that only the sample size n was fixed by the design (of the survey), determine whether they can be considered to be independent.

   (b) How does this result compare with that of a chi-squared test of independence?

2. Recall the data given in TABLE 3 of Chapter 4 of the course notes, reproduced below, concerning a randomized clinical trial of an influenza vaccine.

	RESPONSE			
TREATMENT	Small	Moderate	Large	**Total**
Placebo	25	8	5	38
Vaccine	6	18	11	35
Total	31	26	16	73

   Patients were randomly allocated to two groups, one group given the new vaccine and the other a placebo. The responses were the level of antibody found in the blood six weeks after vaccination. The numbers of patients who were selected to receive the vaccine and the placebo were fixed by the design.

   (a) Is it possible to test for
      (i) independence between the classifying factors, or
      (ii) homogeneity?

      Give reasons for your choice.

   (b) Using R, determine whether the vaccine makes any difference to the response levels.

In Chapter 5, we discussed the extension of log-linear models to three-way tables of counts and higher, including discussion of the following example from Chapter 4. You also have the extended worked example, Example 7, in the handout 'Generalized Linear Models: Worked Examples/Case Studies'. You should explore both these data sets for yourself, looking for an appropriately fitting model for the counts, and considering what this tells us about the relationships between the variables. As a start consider the example given, and analysed, in the lecture notes.

3. TABLE 4: Bishop (1969). Classifies 715 babies according to their 'survival' (survived or died), clinic attended (clinic A or B), and level of ante-natal care received (low or high).

		Survival		
Clinic	Antenatal Care	Survived	Died	**Total**
A	Low	176	3	179
	High	293	4	297
B	Low	197	17	214
	High	23	2	25
	Total	689	26	715

Assuming that only the sample size, 715, has been fixed, so that any model must include a constant term (!), try to identify any association between any of the three response variables, clinic, care and survival.

(i) Verify that the best fitting model has the main effects of each of the variables survival, clinic and care, and the two-way interactions between clinic and care, and clinic and survival.

(ii) How would you interpret this model?

4. Now work through the case study (Example 7, in the handout 'Generalized Linear Models: Worked Examples/Case Studies') at your leisure.