Assignment 6 — Due Dec 6, 2018

1. `online submission` The following data were collected in a study of the relation between parental socioeconomic status (SES) and the mental health of children. Any systematic variation is of interest. Analyze the data and write a brief report (1 page at most).

Parent's	Mental health status			
SES	Well	Mild	Moderate	Impaired
A(high)	64	94	58	46
B	57	94	54	40
C	57	105	65	60
D	72	141	77	94
E	36	97	54	78
F(low)	21	71	54	71

2. In the Bradley-Terry model for ranking k competitors, parameters $\theta_1, ..., \theta_k$ representing 'abilities' are introduced in such a way that the probability π_{ij} that competitor i beats j is a function of the difference in their abilities. In the logit model, we have

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \theta_i - \theta_j.$$

Suppose that seven teams compete in a round-robin tournament, with each pair competing at least once.

(a) Write out the 21×7 model matrix $\mathbf{X}$ for the Bradley-Terry model. (Hint: $21 = \binom{7}{2}$ is the total number of pairs for 7 competitors.)

The data below give the home-team win/loss record of seven American-League baseball teams (Milwaukee, Detroit, Toronto, New York, Boston, Cleveland and Baltimore) in the 1987 season. (New York beat Toronto in two of seven games at NY, and in four of six games at Toronto.)

Home team won/lost record in 1987

	Away Team						
Home Team	Milwaukee	Detroit	Toronto	New York	Boston	Cleveland	Baltimore
Milwaukee	—	4-3	4-2	4-3	6-1	4-2	6-0
Detroit	3-3	—	4-2	4-3	6-0	6-1	4-3
Toronto	2-5	4-3	—	2-4	4-3	4-2	6-0
New York	3-3	5-1	2-5	—	4-3	4-2	6-1
Boston	5-1	2-5	3-3	4-2	—	5-2	6-0
Cleveland	2-5	3-3	3-4	4-3	4-2	—	2-4
Baltimore	2-5	1-5	1-6	2-4	1-6	3-4	—

(b) Fit the Bradley Terry model to these data to obtain a ranking of the teams. Extend this model by including a home-team advantage effect (equal for all teams). Obtain the likelihood-ratio statistic. Comment briefly on the magnitude of home-field advantage.

(c) Estimate the probability that Detroit beats Boston (i) at Boston, (ii) at Detroit, and (iii) on neutral territory.

(d) Does the extended model fit the data? Comment briefly on any patterns in the residuals.

3. Table 1 gives the mean number of children born per woman, the women being classified by place, education, and years since first marriage. Any systematic variation in the number of children is of interest. See `fiji.txt` for the dataset.

Table 1: Mean number of children born to women in Fiji of Indian race, by marital duration, type of place and education. Observed mean values and sample sizes.

Years since first marriage	Type of place: Urban, Education (1)	(2)	(3)	(4)	Type of place: Rural, Education (1)	(2)	(3)	(4)
<5	1.17	0.85	1.05	0.69	0.97	0.96	0.97	0.74
	12	27	39	51	62	102	107	47
5–9	2.54	2.65	2.68	2.29	2.44	2.71	2.47	2.24
	13	37	44	21	70	117	81	21
10–14	4.17	3.33	3.62	3.33	4.14	4.14	3.94	3.33
	18	43	29	15	88	132	50	9
15–19	4.70	5.36	4.60	3.80	5.06	5.59	4.50	2.00
	23	42	20	15	114	86	30	1
20–24	5.36	5.88	5.00	5.33	6.46	6.34	5.74	2.50
	22	25	13	3	117	68	23	2
25+	6.52	7.51	7.54	—	7.48	7.81	5.80	—
	46	45	13	0	195	59	10	0

Education categories are: (1) none, (2) lower elementary, (3) upper elementary, (4) secondary or higher. Lower figures give the number of women in the sample.

(a) Fit an appropriate model describing how the number of children varies with marital age, mother's abode and education. Give a brief synopsis of the arguments justifying your formulation and choice of model, including checks for model adequacy.

(b) Explain the meaning of all parameters in your model. Comment on the major factors affecting fertility.

(c) Construct a 95% confidence interval for the mean number of children born to an urban woman with upper elementary education after ten years of marriage.

(d) Estimate the lifetime average number of children born to rural women with secondary education. Give 90% confidence limits.

4. The file `byss.txt` contains information, obtained from a survey conducted by a large textile company, on the prevalence of byssinosis, a lung disease to which cotton workers are subject. The file lists the observed prevalence of byssinosis (affected, not affected), by race (white = 1; non white = 2), sex (male = 1; female = 2), smoking habits (two levels), length of employment (three levels), and dustiness of the work environment (three levels). In the last three cases, higher-numbered categories denote larger values (more smoking, longer employment and increased dustiness). Parts (a) and (b) are based on the assumption that the main-effects linear logistic model is substantially correct.

(a) Fit the main-effects linear logistic model. Explain how the residual degrees of freedom is calculated for the deviance.

(b) Interpret the coefficient of `sex(2)`. Construct an approximate 90% confidence interval for the odds ratio (males vs females) of contracting byssinosis.

(c) Drop the least significant factor from the model, proceeding until all the remaining factors are significant at the 5% level. Interpret the reduced model thus obtained.

(d) Beginning with the complete main-effects model, look for significant interactions by fitting each of the ten models main effects + one interaction. In judging the significance of interactions, you should bear in mind, at least informally, the effects of selection. After detecting the significant interactions, remove insignificant main effects as described in (c), except for those that are included in interactions. Interpret the model thus obtained.

(e) You are required to write a short report giving details of the excess risk associated with cotton dust. How fast does the risk increase with dust level? If necessary, give separate figures for males and females or for smokers and non-smokers.

(f) Does this analysis suggest that the aetiology of byssinosis is related to sex or race? Explain.