THE CHINESE UNIVERSITY OF HONG KONG
Department of Statistics

STAT4006: Categorical Data Analysis
Problem Sheet 4

The deadline for this Problem Sheet is 5.30pm on Monday 14th December. Please submit your solutions via the link provided on the course Blackboard page - if you must submit your solutions in hard copy, please contact me at jawright@sta.cuhk.edu.hk in advance. **No late submissions will be accepted. A late submission will receive a mark of zero.** Students may discuss set problems with others, but their final submissions must be their own work.

Please answer the following problems. Questions should be answered using a pen, paper, calculator (good practice for your midterm and final). That said, you may use any software you like to find percentiles (i.e. for finding p-values). Show your working.

1. **(Exercise 8.1 from Agresti (2013))** For Table 1, let Y = belief in existence of heaven, x_1 = gender (1= females, 0 = males), and x_2 = race (1=blacks, 0=whites). Table 2 shows the fit of the model

$$\log(\pi_j/\pi_3) = \alpha_j + \beta_j^G x_1 + \beta_j^R x_2, j = 1, 2$$

with *s.e.* in parentheses.

(a) Find the prediction equation for $\log(\pi_1/\pi_2)$.

(b) Using the "yes" and "no" response categories, interpret the conditional gender effect using a 95% confidence interval for an odds ratio.

(c) Find $\hat{\pi}_1 = \hat{P}(Y = \text{ yes })$ for white females.

(d) Without calculating estimated probabilities, explain why the intercept estimates indicate that for white males, $\hat{\pi}_1 > \hat{\pi}_2 > \hat{\pi}_3$. Use the intercept and gender estimates to show that the same ordering applies for black females.

(e) Without calculating estimated probabilities, explain why the estimates in the gender row indicate that $\hat{\pi}_1$ is higher for females than for males, for each race.

(f) For this fit, $G^2 = 0.69$. Deleting the gender effect, $G^2 = 46.74$. Conduct a likelihood-ratio test of whether opinion is independent of gender, given race. Interpret.

		Belief in Heaven		
Race	Gender	Yes	Unsure	No
Black	Female	88	16	2
	Male	54	17	5
White	Female	397	141	24
	Male	235	189	39

Table 1: Belief in the Existence of Heaven Data

	Belief Categories for Logit	
Parameter	Yes/No	Unsure/No
Intercept	1.785 (0.168)	1.554 (0.172)
Gender	1.044 (0.259)	0.254 (0.269)
Race	0.703 (0.411)	-0.106 (0.438)

Table 2: Heaven Data - Fitted Values

2. **(Exercise 8.37 from Agresti (2013))** Consider the logit model,

$$\text{logit}[P(Y \leq j)] = \alpha_j + \beta_j x,$$

not having proportional odds form.

   (a) With continuous x taking values in $(-\infty, \infty)$, show that the model is improper in that cumulative probabilities are misordered for a range of x values.

   (b) When x is a binary indicator, explain why the model is proper but requires constraints on $(\alpha_j + \beta_j)$ (as well as the usual ordering constraint on $\{\alpha_j\}$) and is then equivalent to the saturated model.

3. There are 9 different hierarchical loglinear models can be fit to a contingency table with three variables X, Y and Z. List all models by using the notation adopted in class notes. For each model you have, state the structure among X, Y and Z.

4. For the saturated model with "Belief in Afterlife" data (Example 10.1 in the lecture notes), Table 3 reports the $\{\lambda_{ij}^{XY}\}$ estimates: Show how to use the data in Table 3 to estimate the odds ratio.

Parameter			*df*	Estimate	Std. Error
gender*belief	females	yes	1	0.1368	0.1705
gender*belief	females	no	0	0.0000	0.0000
gender*belief	males	yes	0	0.0000	0.0000
gender*belief	males	no	0	0.0000	0.0000

Table 3: Afterlife Data

5. In a survey study, 2,276 students are asked whether they had ever used alcohol (A), cigarettes (C), or marijuana (M) in their final year of high school in a non-urban area near Dayton, Ohio. The fitted values for several loglinear models are shown in Table 10.6 from the Chapter 6 notes.

   (a) Use AIC and BIC to select the best model based on the information given in Table 4.

Model	G^2	*df*
(A,C,M)	1286.0	4
(AC,M)	843.8	3
(AM,C)	939.6	3
(CM,A)	534.2	3
(AM,CM)	187.8	2
(AC,AM)	497.4	2
(AC,CM)	92.0	2
(AC,AM,CM)	40.4	1

Table 4: Model Selection

   (b) Write down the loglinear model you identified in (b). Also show clearly how can you derive the corresponding logit model, regarding the whether they had ever used cigarettes (C) as the response variable.

6. The 1988 General Social Survey compiled by the National Opinion Research Center asked: "Do you support or oppose the following measures to deal with AIDS? (1) Have the government pay all of the health care costs of AIDS patients; (2) Develop a government information program to promote safe sex practices, such as the use of condoms." Table 5 summarizes fits of loglinear models about health care costs (H) and the information program (I), classified also by the respondent's gender (G).

   (a) Explain why these model (GH, GI, HI) has one degree of freedom.

   (b) Use Table 5 to test which interaction terms are significant using likelihood ratio tests. Which models would you like to fit next?

THE END

Model	df	Deviances	p-value
(GH, GI)	2	11.67	0.0029
(GH, HI)	2	4.127	0.1270
(GI, HI)	2	2.383	0.3038
(GH, GI, HI)	1	0.3007	0.5834

Table 5: AIDS Survey Model Fits