STA4030: Categorical Data Analysis

Assignment 4

Due Date and Time: December 18, 2020 (Friday), 10:00PM

INSTRUCTION:

- Please scan your answers in **one single .pdf file** and submit via Blackboard System.
- **Late submissions** will receive a mark of zero.
- Students may discuss set problems with others, but your final submissions must be your own work.
- All these questions should be answered using a pen, paper, calculator (good practice for your midterm and final).
- You may use any software you like, e.g., R, Python, Excel, etc., to find the percentiles regarding relative distributions (for example, to find p-values).
- Show and write down your solutions in detail and clearly.

Problem Set 4:

1. (Exercise 8.1 from Agresti (2013)) In Table 1, 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 SE (standard error) reported in the parentheses.

		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.

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

	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.

(b). Use the "yes" and "no" response categories to interpret the conditional gender effect by using a 95% confidence interval for the 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, $G^2 = 47.64$. Conduct a likelihood ratio test of whether opinion is independent of gender, given race. Try to interpret the results.

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

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

(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 introduced in lecture notes. For each model, try to state the structure connection among X, Y and Z.

4. 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 lecture slides 11 (refer to Page 25 and Page 26 of lecture slides 11).

(a). Use AIC and BIC to select the best model based on the information given in Table 3.

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)	0.4	1

Table 3: Model Selection.

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

5. 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 4 summarizes fits of loglinear models about health care costs (H) and the information program (I), classified also by the respondent's gender (G).

Model	*df*	Deviance	*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 4: AIDS Survey Model Fits.

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

(b). Use Table 4 to test which interaction terms are significant using the likelihood ratio tests. Which models would you like to fit next?

THE END