National University of Singapore
Department of Statistics and Applied Probability
ST5213 Categorial Data Analysis II
Assignment 2

Important: This assignment accounts for 20% of your final grade. Your work for this assignment must be uploaded onto the LumNUS submission folder by **7 pm 5 April 2021 (Monday)**. Any student failing to submit work by the deadline will receive a penalty for late submission unless the lecturer is advised as soon as possible of any extenuating circumstances. Please upload a **single pdf file** labeled using your student number, e.g., A0123456M.pdf.

Plagiarism: The work that you submit must be your sole effort (i.e. not copied from anyone else). You may be severely penalized If found guilty of plagiarism.

The two assignment tasks involve the analysis of some data and you should submit a formal report for each task.

Format: The report for **each task** should not exceed **two pages** of A4 paper including any relevant figures or tables. The first 4 pages should consist of your reports for the two tasks. `R` code and output should be attached as appendix **after** these four pages. Please write down your name and student number at the top of the first page. Do not include any cover page. Please use Times New Roman font (11-12 point). You may be penalized for exceeding the page limit or if any of the above instructions are not followed.

The aim of the report is to convey the methodology and results of your data analysis in a clear and concise manner with appropriate use of figures or tables for summarization. Marks will be awarded for

- Exposition: your report should be well-organized. You should aim to write in a clear and concise manner.
- Statistical content: marks will be awarded for the correct use of appropriate statistical techniques, and for the correct interpretation of results from these techniques.
- Completeness: ensure that you have answered all the parts of each question.

Note: All results must be included in the report itself and not in the appendix. Whenever you are asked to fit a model, you should write down the equation of the fitted model with the estimated coefficients. You will be assessed based on the report alone. I will refer to your R code and output only if I wish to locate any source of error.

Task 1. A cohort study in South Africa is designed to follow children born between April and June 1990 in hopes of identifying risk factors for cardiovascular disease. After five years, the children were invited to participate in interviews. Many children did not participate in these interviews, leaving open the possibility of biases if inferences were made based on those who participated. Morrell (1999) gave data comparing the children who participated to those who did not with respect to whether the mother had medical aid at the time of birth.

(a) Based on the information in Table 1, use Pearson's Chi-square test to determine if there is evidence of a relationship between medical aid status and participation in the interviews. Compute the marginal odds ratio between medical aid status and participation in the interviews and interpret.

	No interview	Interview
Had medical aid	195	46
No medical aid	979	370

Table 1

(b) The study further classified children by their racial group, with the results given in Tables 2 and 3. Use Fisher's exact test to determine if there is evidence of a relationship between medical aid status and participation in the interviews for (i) white children and (ii) black children. Explain how the probabilities of the observed tables and p-values are computed in Fisher's exact test. Compute the conditional odds ratios between medical aid status and participation in the interviews given race and interpret.

White children

	No interview	Interview
Had medical aid	104	10
No medical aid	22	2

Table 2

Black children

	No interview	Interview
Had medical aid	91	36
No medical aid	957	368

Table 3

(c) Explain why the marginal association is so different from the conditional associations.

Task 2. The following are data on smoking from a survey of seventh graders (age: 1 = 12 or younger, 2 =13 or older):

				Smoking	
Family structure	Race	Gender	Age	None	Some
Both parents	Black	Male	1	27	2
			2	12	2
		Female	1	23	4
			2	7	1
	White	Male	1	394	32
			2	142	19
		Female	1	421	38
			2	94	11
Mother only	Black	Male	1	18	1
			2	13	1
		Female	1	24	0
			2	4	3
	White	Male	1	48	6
			2	25	4
		Female	1	55	15
			2	13	4

Search for the loglinear model that can best explain the association patterns in the contingency table by treating

(a) smoking and family structure as response variables and the rest as explanatory variables.

(b) smoking as a response variable and the rest as explanatory variables.

In each case,

i. State the minimal model.

ii. Give the symbol for the loglinear model that best describes the data and explain how it was built. Write down the coefficients of the fitted model.

iii. Represent the conditional independence structure in the loglinear model using an association graph and explain whether it is a graphical model. Give the symbol of another model that has the same association graph, but is not a graphical model.

iv. Interpret the associations in the loglinear model, taking into account conditional independence, collapsibility and odds ratios.

v. Explain whether the zero cell in the contingency table affects your analysis.

vi. State the logit model equivalent to the selected loglinear model for (b).