National University of Singapore
Department of Statistics and Applied Probability
ST5213 Categorial Data Analysis II
Assignment 1

Important: This assignment accounts for **20%** of your final grade. Your work for this assignment must be uploaded onto the LumiNUS submission folder by **7 pm 8 March 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. The file `menarche.txt` from LumiNUS refers to a survey of teenage girls in Warsaw in 1965 (Milicer and Szczotka, 1966). The variables recorded are:

Variable	Description
`Age`	The average age of girls in each sample group. The groups were school classes and hence approximately age-homogeneous.
`Total`	The number of girls in each sample group.
`Menarche`	The number of girls in the group who have reached menarche.

(a) Construct plots of the (1) sample proportions against age, (2) logit of the sample proportions against age and (3) probit of the sample proportions against age. Describe your observations of these plots.

(b) Build a model that best describes the relationship between age and the probability that a girl has reached menarche by considering both logit and probit links, and higher order terms in age.

Provide a step-by-step description of how model comparison is performed and summarize your results using tables. Conclusions must be supported with evidence.

(c) Fit a logistic regression model using $\log(\text{Age}-9)$ and $\log(18-\text{Age})$ as predictors. Is this model better than the one selected in (b)? Discuss the advantages and disadvantages of using this model for prediction as compared to the one selected in (b).

(d) Fit the complementary log-log model to these data with a cubic polynomial in Age. For this model, compute estimates of the decile ages, that is, the ages at which the probability that a girl has reached the age of menarche is $0.1, 0.2, \ldots, 0.9$, and their standard errors. Show your working and derivation clearly.

Task 2. Fowlkes et al. (1988) reported the results of a survey of employees of a large national corporation to determine how job satisfaction depends on the demographic variables: race, gender, age and the regional location of the local company. The data is available in the file `satisfaction.csv` from LumiNUS. The variables recorded are

Variable	Description
`Satisfied`	The number of employees whose response was "satisfied" at each setting.
`Notsatisfied`	The number of employees whose response was "not satisfied" at each setting.
`Gender`	female (`F`) or male (`M`).
`Race`	White (`W`) or other (`O`). The "other" category includes black, Hispanic and other minorities.
`Age`	less than 35 age group (`<35`), 35–44 age group (`35-44`) or greater than 44 age group (`>44`).
`Region`	seven levels; Northeast (`NE`), Mid-Atlantic (`MA`), Southern (`S`), Midwest (`MW`), Northwest (`NW`), Southwest (`SW`), Pacific (`P`).

(a) Build a logistic model that best describes how employee satisfaction is related to the four demographic variables. Provide a step-by-step description of how the model is built and support your conclusions with evidence.

(b) Fit the logistic model (Region + Race + Gender*Age) and interpret all the parameter estimates.

(c) Fit the probit model (Region + Gender*Race + Gender*Age) and use it to compute a 95% confidence interval for the probability of satisfaction for a female white employee aged 35-44 working in the Pacific Region.

(d) Perform an informal goodness of fit test for the probit model in (c). Explain whether there is any overdispersion. Revise the confidence interval found in (c) using the quasi-likelihood approach to account for any overdispersion.