Project Outline and Expectations

STSCI/BTRY 4110/5160 Categorical Data Analysis, Fall 2023 Prelim 2 Take Home Project

Your project report should include the following five sections:

- Executive Summary
  - Research question is clear, results are summarized, conclusion is present.
  - Indicate important variables and the direction of their effects.
- Introduction
  - Brief information about why the research question is important
  - Description of overall analysis strategy
- Description of Subjects
  - How data were checked and cleaned
  - Explanation of how you dealt with missing values
  - Description of characteristics of subjects (percentages for categorical variables and outcome variable, medians/stdev/min/max for numerical variables). Create table with summary. Also include verbal description of these characteristics.
  - In the table you should have two columns related to the outcome variable. See example.
- Results
  - Two-way contingency analysis of variables (one at a time) and their association with outcome variable
  - Include mosaic plots for each categorical variable with outcome variable
  - Exploration and choice of the best transformation for every significant, numerical explanatory variable (include at least one "slicing-dicing"plot of empirical log-odds)
  - Consideration of whether levels of a categorical variable warrant being combined. Justify.
  - Multivariable analysis using logistic model
    - Selection of significant covariates
    - Description of the process in fitting the best model. (Don't need to present all hypotheses tests done or the specific details of each step. Describe the overall methods used and steps taken.)
      - Consideration of interactions (no more than two-way)
    - Final model is written in general form, with description of coding for each variable in the model.
    - Table is created with parameter estimates, odds ratios, odds ratio confidence intervals, p-value for each variable in the final model
    - Assessment of the overall goodness-of-fit of the model (Show all three: classification table, goodness-of-fit test and ROC curve)
  - Success probabilities of representative sub-populations should be described, with tables and figures, as appropriate. Here is a detailed description of what you should produce for a probability plot: Choose two variables that are significant in the model. One ("A") should be continuous/numerical and the other ("B") should be categorical. Calculate success probabilities over the full range of both variables, holding values of all other variables at their mean or median or most common values. Create a plot. The x-axis should show the range of values of A, the y-axis will be success probability, and there should be 2 or more lines, each corresponding to a different level of the B variable. At a minimum, you should produce two probability plots as described above.
- Discussion
  - Clearly states conclusions about the effect and direction of each variable in the final model
  - Discussion of sleep duration on the incidence of cardiac events
  - Discussion of problems encountered during the analysis (if any)
  - Possible avenues for further research on the subject are mentioned

Overall Writing and Style

- Correct grammar and spelling
- Data/figures
  - Meaningful units are used and included
  - Graphs are of the appropriate type and are labeled accurately
- Paper is logically organized and easy to follow and easy to read
- Paper is concise – unnecessary or irrelevant information is not added

Page Limit for Report: 12 pages (cover sheet does not count towards this limit)
R Code: Submit in a different file than the report.