Psych 6136: Project 2

Instructions

Several research problems, involving categorical data methods--- multi-way tables, loglinear models, logistic regression or correspondence analysis are described below.

For TWO of these problems,

- Carry out appropriate analyses, guided by (but not limited to) the questions or suggestions posed. Feel free to make up your own questions.
- Create meaningful and useful displays to explore the data and explain the results
- Write up a brief research report including (a) a problem description, (b) methods of analysis and (c) results section, and (d) a summary/discussion/conclusions section. If you wish, you can include some of the details from your analyses in the previous steps as an appendix, "Supplementary materials".

You can use any statistical or other software you like, though you may find that some of the steps or questions are easier to do in R. [You can use `write.csv()` to export an R data set in a form you can import into other software.]

You should submit your report both by email (PDF or MS Word) and in hardcopy. If you used R, please also submit the .R or .Rmd script(s) you used for your analyses by email. Please name these files along the lines of 'YourName-Project2'.

Problems

1. **Dayton Survey data**: The data set DaytonSurvey in vcdExtra gives a 2 x 2 x 2 x 2 x 2 table in the form of a frequency data frame containing results of a survey of high school students regarding whether they had every used alcohol (A), cigarettes (C) or marijuana (M), classified by sex (S) and race (R). The goal here is first to understand the associations among the variables A, C, M, and then determine whether and how they differ with sex and race.
    a. For the first goal, you can be guided by the questions posed in Exercises 9.1 and 9.2 in the revised version of VCDR. These ignore (collapse over) race and sex, and ask you to fit and plot various loglinear models.
    b. For the second goal, extend the analysis to include the variables sex and race as potential explanatory variables of substance use. There are several ways to go about this: You could fit logit models predicting the use of each substance from the remaining variables, or loglinear models that include the sex*race association.
    c. Use analysis and plots of log odds ratios for the association between A and M in relation to the remaining variables.

2. **Accident data**: The data set `Accident` in vcdExtra gives a 5 x 2 x 4 x 2 table of data on accidents in France representing the combinations of age (5 levels), result (died or injured), mode of transportation and gender. Questions: How are these variables associated? What factors determine whether an accident results in a fatality?
    a. Use loglm() or glm() to fit the model of mutual independence, Freq ~ age + mode + gender + result to this data. (Technical note: age is contained as an ordered factor in the data set, and glm() treats such variables differently than ordinary factors.)
    b. Use mosaic() to produce an interpretable mosaic plot of the associations among all variables under the model of mutual independence. Try different orders of the variables in the mosaic. (Hint: the `abbreviate` component of the `labeling_args` argument to mosaic() will be useful.)
    c. Use some of the methods of Correspondence Analysis (Ch. 6) to investigate these associations further. An MCA is probably the most straight-forward, but the stacking approach might also be useful.
    d. Treat `result` (Died" vs. Injured") as the response variable, and either a loglinear model or a logistic regression/logit model to predict a fatality. Consider the possibility that there may be interactions among the predictors. Make some useful plots to illustrate or support your conclusions.
3. **Housing data**: The data set `housing` in the MASS package gives a 3 x 3 x 4 x 2 table relating satisfaction (Sat) of residents in Copenhagen with their housing to their perceived degree of influence (Infl) on management of the property, the Type of apartment and the degree of contact (Cont) residents have with other residents. Question: How does satisfaction vary with these factors?
    a. Exercise 7.6 in VCDR (Ex. 8.2 in the revised version) describes some analyses and steps to investigate this question.
4. As a final option, you can use the one question **you did not attempt** from Project 1.