Psych 6136: Project 1

Instructions

Several research problems, involving categorical data methods--- two-way and multi-way tables, loglinear models, correspondence analysis or logistic regression are described below. Your implicit goal is to convey to me your understanding of the methods of the course and how you would use them in practice, e.g., in a research report

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. You can include the steps and results for the main analyses inline if this serves your narrative, or include some of the details from your analyses in the previous steps as an appendix to the problem, "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 work 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. [For R markdown, you might use the template linked on the course web page.]

Problems

1. **Minnesota High School Graduates**: The `Hoyt` data in vcdExtra gives a 4 x 3 x 7 x 2 table classifying nearly 14000 graduates by (a) post-high school `Status`, (b) `Rank` in their graduating class, (c) father's `Occupation`al status (7 levels, from 1=High to 7=Low) and (d) `Sex`. Status is considered the outcome variable. How does it depend on the other factors?
    a. Analyze and display the associations between status and each of the other variables individually. In doing this, you might also consider focusing attention on the distinction between College and Non-college Status.
    b. Consider and analyze the associations among the predictor variables: Rank, Occupation and Sex. Is there anything of interest worth mentioning or analyzing in further detail? Is it useful to consider collapsing some of the categories?

c. Start by fitting a simple model of joint independence of Status from the predictors, then add additional associations as you consider necessary until you obtain a reasonably well-fitting model.

2. **Student Opinion about the Vietnam War**: The data set `Vietnam` in vcdExtra gives a 2 x 5 x 4 contingency table in frequency form reflecting a survey of student opinion on the Vietnam War at the University of North Carolina in May 1967. The table variables are `sex`, `year` in school and `response`, which has categories: (A) Defeat North Vietnam by widespread bombing and land invasion; (B) Maintain the present policy; (C) De-escalate military activity, stop bombing and begin negotiations; (D) Withdraw military forces Immediately. How does the chosen response vary with sex and year?
    a. Construct some useful plots showing the proportions of each response category by sex and year of study. It may be helpful to add some smoothed or fitted curves to aid interpretation.
    b. Analyze the associations between response and each of sex and year separately, using graphs and statistical tests. Note that both year and response can be considered ordinal factors.
    c. Fit loglinear and/or polytomous logit models designed explain the relationship between response and year of study.

3. **Accident data**: The data set `Accident` in vcdExtra gives a 4-way table of frequencies of traffic accident victims in France in 1958. It is a frequency data frame representing a 5 × 2 × 4 × 2 table of the variables age, result (died or injured), mode of transportation, and gender. What factors determine mortality in a traffic accident?
    a. Use `loglm()` to fit the model of mutual independence, Freq ~ age+mode+gender+result to this data set.
    b. Follow this with `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 to avoid some overlap of the category labels.)
    c. Now, carry out a multiple correspondence analysis of these data. You will need to convert the data frame to a 4-way table,
    ```
    accident.tab <- xtabs(Freq ~ age + mode + gender + result, data = Accident)
    ```
    What can you say about the degree to which the table can be approximated by a 2D solution?
    d. Construct an informative 2D plot of the solution, and interpret in terms of how the variable result varies in relation to the other factors. [Hint: `vcdExtra::mcaplot` may be helpful.]

4. **Risk factors for low infant birth weight**: The data set `birthwt` in the MASS package gives data on 189 babies born at Baystate Medical Center, Springfield, MA during 1986. The quantitative response is `bwt` (birth weight in grams), and this is also recorded as `low`, a binary variable corresponding to `bwt < 2500` (2.5 Kg). The goal is to study how this varies with the available predictor variables.

The variables are all recorded as numeric, so in R it may be helpful to convert some of these into factors and possibly collapse some low frequency categories. The code below is just an example of how you might do this for some variables.

```
birthwt2 <- within(birthwt, {
    race <- factor(race, labels = c("white", "black", "other"))
    ptd <- factor(ptl > 0)  # premature labors
    ftv <- factor(ftv)      # physician visits
    levels(ftv)[-(1:2)] <- "2+"
    smoke <- factor(smoke>0)
    ht <- factor(ht>0)
    ui <- factor(ui>0)
  })
```

a. Make some exploratory plots showing how low birth weight varies with each of the available predictors. In some cases, it will probably be helpful to add some sort of smoothed summary curves or lines.
b. Fit several logistic regression models predicting low birth weight from these predictors, with the goal of explaining this phenomenon adequately, yet simply.
c. Use some graphical displays to convey your findings.
d. [optional] Because actual birth weight is available there are other approaches you could explore. One would be to start with a classical regression of bwt on the predictors you used in your categorical model, and then compare results or the categorized predicted values, fitted(bwt) < 2500 with the results of your logistic model.